DOCUMENT RESUME

ED 434 535 FL 026 014

AUTHOR Kang, Sungwoo
TITLE Comparing the Effects of Native Language and Learning Experience on the University of Illinois at Urbana-Champaign English as a Second Language Test.
PUB DATE 1997-05-00
NOTE 121p.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS College Students; *Educational Background; *English (Second Language); Higher Education; *Language Tests; Languages; Student Placement; Testing
IDENTIFIERS University of Illinois Urbana Champaign

ABSTRACT

A study investigated the relationship between two factors, native language and language learning experience, on university students' performance on the English Placement Test used at the University of Illinois at Urbana-Champaign. Subjects were 203 students who took the test in August 1994 and completed a questionnaire about their learning experiences. The test consisted of three sections: language structure; video-essay; and pronunciation. Statistical analysis of test results and student information revealed that native language was more influential on test performance than learning experience. Language groups differed in both the pronunciation and video-essay sections, whereas learning-experience groupings differed only in the video-essay section. Neither grouping showed statistically significant differences on the structure section, but native language groupings did perform differentially, although not to statistical significance. Results suggest that test-takers' native languages should be given more weight than learning experience in placing students into appropriate English-as-a-Second-Language classes if variables other than English language proficiency are considered in placement. Appendices contain an English Placement Test information form and simultaneous item bias test output. Contains 48 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 434 535

Comparing the effects of native language and learning experience on the University of Illinois at Urbana-Champaign English as a Second Language Test

Sunghee Kang

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Sungwoo Kang

May, 1997

University of Illinois at Urbana-Champaign

BEST COPY AVAILABLE

BEST COPY AVAILABLE

FL026014

ABSTRACT

Test performance and language proficiency have been found to be affected by various test-taker characteristics such as native language, learning style, personality, and instruction type. Understanding the effects of these characteristics may help diagnosis and explain problems and strengths of a learner or a group. This study investigated the relative effects of two test-taker grouping characteristics, (a) native language and (b) language learning experience, on test performance on the University of Illinois at Urbana-Champaign English Placement Test (UIUC EPT) to provide a better understanding of test performance, and to provide a basis for the use and interpretation of the UIUC EPT.

The subjects were 203 students who took the UIUC EPT during August 1994 and who completed a questionnaire about their learning experiences. Based on the information from the questionnaire, they were grouped according to their native language and learning experiences.

The UIUC EPT consisted of three sections; structure, video-essay, and pronunciation. To compare the relative effects of native language and learning experience, a two-way ANOVA was used for the structure section and a chi-square statistic was used for both the video-essay section and the pronunciation section. For the structure section, how the items functioned differentially according to the group

membership was studied by using the simultaneous item bias test (SIB test).

Native language appeared more influential on performance on the UIUC EPT than learning experience. Grouping by both native language and learning experience did not result in group differences in the structure session of the exam. Language groups were different in both the pronunciation section and the video-essay section whereas the learning experience groups differed only in the video-essay section. Even though neither grouping resulted in group differences on the structure section, the item level comparison showed that test takers performed more differentially when they were grouped according to native language. The collective amount of differential functioning appeared significant only when the different native language groups were compared, but not when the performance was compared among the different learning experience groups.

These findings suggest that test takers' native languages be given more weight than learning experience in placing test takers into appropriate ESL classes if variables other than English proficiency are included in the EPT administration.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my adviser, Professor Fred Davidson, and to my committee members, Professor Gary Cziko and Katherine Ryan. Without their professional and academic guidance, this thesis would not be possible.

I thank Dr. Davidson for his advice, valuable comments, and encouragement.

As to Dr. Cziko, I am grateful for helping me with the problems that I have struggled with.

I appreciate Dr. Ryan's guidance and suggestions concerning statistics and measurement issues.

I would also like to thank all students who returned the questionnaire.

TABLE OF CONTENTS

CHAPTER I	INTRODUCTION	1
CHAPTER II	LITERATURE REVIEW	5
2.1.	Changes in Conceptions of Validity . . .	5
2.2.	Native Language, Instructional Method, and Second Language Proficiency	8
2.2.1.	Native language	9
2.2.2.	Native country and language instruction .	17
2.3.	Implication for Research	24
CHAPTER III	BACKGROUND INFORMATION AND RESEARCH QUESTION	28
3.1.	UIUC EPT	28
3.2.	Native Language	29
3.3.	Language Learning Experience	30
3.4.	Research Questions	32
CHAPTER IV	METHOD	35
4.1.	Subjects	35
4.2.	Grouping	35
4.3.	Analyses	39
CHAPTER V	RESULTS	42
5.1.	Group Differences in Each Section	42
5.1.1.	Structure	42
5.1.2.	Pronunciation	44
5.1.3.	Video-Essay	46
5.2.	Dimensionality and DIF in the Structure Section	48

TABLE OF CONTENTS (CONTINUED)

CHAPTER VI. DISCUSSION 53

CHAPTER VII CONCLUSION 62

REFERENCES 65

APPENDICES

A. EPT INFORMATION FORM 72

B. SIB TEST OUTPUT 74

1. Grammar-and-Reading vs. Controlled-Oral-Language 74

2. Grammar-and-Reading vs. Communication . . . 80

3. Controlled-Oral-Language vs. Communication. 87

4. Altaic vs. Indo-European 93

5. Indo-European vs. Sino-Tibetan 100

6. Sino-Tibetan vs. Altaic 106

LIST OF TABLES

Table 3.1.	Classification of language families	30
Table 4.1.	The summary of the result of examinee's learning experience judgment	37
Table 4.2.	Language families	38
Table 4.3.	Number of examinees by language families and instructional methods	39
Table 5.1.	Summary descriptive statistics of structure scores by language family and learning experience	43
Table 5.2.	Descriptive statistics by language family .	43
Table 5.3.	Descriptive statistics by learning experience	43
Table 5.4.	ANOVA result of the structure section . .	44
Table 5.5.	Chi-square analysis on the pronunciation section	45
Table 5.6.	Cell chi-square analysis for the pronunciation section when the examinees were grouped by language family	46
Table 5.7.	Chi-square analyses on the Video-Essay section	47
Table 5.8.	Cell chi-square analysis when the examinees were grouped by language family	47
Table 5.9.	Cell chi-square analysis when the examinees were grouped by learning experience . . .	48
Table 5.10.	DIMTEST statistic	49
Table 5.11.	Comparisons between language family groups	52
Table 5.12.	Comparisons between learning experience groups	52
Table 6.1.	Relative language proficiency profiles of three language groups	53
Table 6.2.	Test of association between native language groups and learning experience groups . .	55
Table 6.3.	Relative language proficiency profiles of three learning experience groups	56

LIST OF TABLES (CONTINUED)

Table 6.4. Sample DIF items when the language groups are compared 58

Table 6.5. Sample DIF items when the learning experience groups are compared 59

LIST OF FIGURE

Figure 4.1. Classification of learning experience 36

CHAPTER I. INTRODUCTION

Each year over 40,000 foreign students (U.S. Bureau of the Census, 1994) come into post-secondary educational institutes in the United States. The number of students who apply to U.S. universities and colleges is even larger. One concern arising from this large number of students is whether they are ready to study in American universities or colleges. Those institutes have to screen the applicants. Foreign applicants are asked to show not only their academic ability, but also enough English proficiency to study in those institutions. Most of the universities and colleges require TOEFL (Test of English as a Foreign Language) scores as proof of English proficiency.

Most universities set minimum limits of TOEFL scores for students to get admitted. However, scores above these limits do not necessarily indicate adequate proficiency for studying in America. The TOEFL scores may not be adequate indicators of the necessary proficiency in all skills. The TOEFL[1] tests only receptive skills, such as reading, listening, and structural knowledge. It does not test productive skills such as writing and speaking. Many universities are administering English Placement Tests (EPTs) to admitted students to assess their English proficiency more thoroughly and help them with their weaknesses. Those universities also provide ESL (English as a Second Language) courses at various levels to

[1] Most universities do not require TSE (Test of Spoken English) or TWE (Test of Written English) for admissions.

help students improve their English. Students are assigned to the appropriate ESL classes based on their EPT scores.

In most universities, placement is made solely based on the EPT results. EPTs usually consist of several sections, such as pronunciation, listening, reading, structure, etc. According to their section scores, students are placed in classes where they can improve their skills. However, EPT scores may not be enough to diagnose learners' language proficiency. EPT takers are from various backgrounds: different nationalities, different native languages, different learning styles and different learning experiences. These characteristics may have affected them to form different views of language and different profiles of language skills (Farhady, 1982). Without any consideration of these characteristics, placing different students into a classroom may make the ESL instruction inefficient (Mitchell, 1991). In an ESL classroom, a teacher has to deal with many different aspects of students' characteristics, and students have to adjust themselves to a new learning environment. However, ESL instruction can be more efficient. If students' problematic areas in learning English are related to their characteristics, that is, students sharing certain characteristics may share the same problems, considering students' characteristics as well as their language proficiency in designing ESL classes may make instruction tuned more to students' needs.

Practically, it is not possible and reasonable to consider all learner characteristics. What one can do is to utilize the one or two most important variables which have been shown to affect second language proficiency development. Many studies have been done to examine the degree to which various characteristics affect students' performance on English proficiency tests. Those characteristics are gender (Farhady, 1982; Hosley, 1978; Ryan & Bachman, 1992), first language (Alderman & Holland, 1981; Dunbar, 1982; Oltman et al., 1988; Ryan & Bachman, 1992; Sawatdirakpong, 1993; Swinton & Powers, 1980;), field of study (Farhady, 1982; Spurling & Ilyin, 1985), native country (Hosley, 1978; Farhady, 1982; Spurling & Ilyin, 1985), and academic status (Farhady, 1982; Spurling & Ilyin, 1985). These studies have showed that those characteristics affect students' performance in one way or another. Among these variables, first language and native country have been two very popular topics and proved to be very influential characteristics on language proficiency.

It is not hard to imagine why these two characteristics are so influential. First language may be the fundamental resource and starting point of second language learning. The pace of second language learning can vary depending on how similar the structure of the second language is to that of the first language (Zobl, 1982). If two languages have a similar structure, the pace of learning the structure may be faster than the pace of learning other structures.

Research on the effects of native country showed that learners' performance varied according to nationality. This may be because learners may have different learning experiences. Due to different educational and social policies and needs of each country, one language skill is emphasized more than the other. Learners may also have different views, conceptions, and perceptions of language (Farhady, 1982). Unfortunately, however, little research has been done to explain the effects of various learning experience. Learners may have different learning experience even though they have the same nationality. To understand test performance better, the effect of learning experience needs direct scholarly attention.

The present study examined and compared the effects of two test taker characteristics: first language and learning experience on the test performance on the UIUC EPT. This study has two objectives. The first objective is to understand test performance on the University of Illinois at Urbana-Champaign English Placement Test (UIUC EPT). The second objective is to obtain some insights for how to interpret the EPT test scores and how to design ESL classes to help students better.

CHAPTER II. LITERATURE REVIEW

The present chapter consists of four sections. The first section discusses how the concept of validity has changed. The new understanding of validity concept provides a framework within which the present study should be understood as a process of construct validation. The second and third sections review the literature on the effects of native language and learning experience, respectively. The last section discusses the implications for the present study.

2.1. Changes in Conceptions of Validity

There have been many changes recently in the concept of validity. Validity, which was broken into several distinct types of validity has evolved as a unitary concept (Messick, 1989). Messick (1989) and Shepard (1993) provided a summary of the changes of the concept of validity by reviewing the Standards for Educational and Psychological Tests and Manuals which have been published four times from 1954 to 1985 by APA, AERA, and NCME. Each edition reflected how validity was regarded and sought at the time of the publication.

Four types of validity were identified in the 1954 Standards: content validity, predictive validity, concurrent validity, and construct validity. Content validity is based on professional judgments about how well the content of the test sample represents the subject matter about which the conclusion is drawn. Construct validity is the degree to which students' performance on a test reflect their true

mental trait. Both concurrent validity and predictive validity indicate the degree to which the test scores estimate students performance on the basis of external criteria. Criteria for concurrent validity are students' present standings on other tests and criteria for predictive validity are students' future or past standings on other tests.

In the following version of the Standards (1966), the last two types of validity pertaining to outside criteria were reduced to one type, criterion-related validity. Inter-relatedness of three types of validity began to be recognized at least in theory in the next revision, the 1974 Standards. It stated "These aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically" (p. 26). Even though inter-relatedness was recognized, many authors still argued that content validity was sufficient to establish the meaning of scores (Shepard, 1993).

In the 1985 edition of the Standards, the distinction between types of validity disappeared. Validity was described as a unitary concept, meaning "appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p. 9). Under the new concept, validation focuses not on a type of validity but on the relation between the evidence and the inference drawn from test scores (Messick, 1989, p. 16). Previously separated categories of validity became referred to as content-related,

criterion-related and construct-related evidence of validity (APA, 1985, p. 9).

Along with the new understanding of validity, construct validity became a cover term for all facets of validity (Messick, 1989; Moriyama, 1994; Shepard, 1993). A construct is a latent trait to be measured or estimated. Therefore, construct validity can be sought from an integration of any evidence pertaining to interpretation or meaning of the test scores. In this sense, construct validity also subsumes content relevance and criterion-relatedness (Messick, 1989).

The fundamental issue in construct validity is the degree to which inferences and actions based on test scores are supported by empirical evidence and theoretical rationales (Messick, 1989). For example, if a test is designed to test language proficiency, inferences drawn from test scores about examinees' language proficiency should reflect the actual language proficiency of the examinees. Low scores on the test should be matched with the low proficiency in the actual language use.

Construct validation is a process of providing evidence for inferences based on test scores. Studying the effects of test taker characteristics on second language proficiency tests can be understood as a basis of seeking construct validity. Language proficiency test takers are usually not homogeneous in their backgrounds: culture, native language, educational background, personality, nationality, reasons for test taking, and so on. These factors can be the sources of

variation among test takers. Dealing with these factors may depend on how language proficiency is defined. Their effects can be subsumed as a part of language proficiency or they may be dealt as the sources of bias (Bachman, 1990). However, no matter how language proficiency is defined, the influences of those variables are present in one's performance. Understanding the effects of these factors may help test users to make correct inferences about test takers' language abilities and diagnose the difficulties test takers have.

As a construct validation process, this study provides a validated basis for correct use and inferences of UIUC EPT scores. This study investigated the effects of two characteristics, native language and learning experience, which have been considered the most important factors both in second language learning and language testing research. This will help to diagnose the problems of test takers and design the appropriate ESL classes which will help students with their problems.

2. 2. Native Language, Instructional Method, and Second Language Proficiency

Researchers in second language learning have come to agree that second language learning is not a simple process of instruction-and-learning but a creative and dynamic process in which learners learn language by constructing hypotheses of language rules and testing them (Brown, 1983; Larsen-Freeman & Long, 1991). The second language learning process does not occur in a uniform way for all learners.

Learners are different in many aspects as well as learning abilities. These characteristics have been found to affect the mastery of second language skills in one way or another; native language (Eckman, 1977; Kellerman, 1977; Zobl, 1982); learning environment (Dulay, Burt, & Krashen, 1982), instruction (Krashen, 1985; Long, 1988; Perkins & Larsen-Freeman, 1975; Wode, 1981), learning style, culture, etc.

Investigating the relative importance of two characteristics will help the test user understand test takers proficiency profiles and make correct placement decisions. The next two sections will review the literature which studied the effects of the two characteristics.

2.2.1. Native language

Among many test taker characteristics, native language has been the most popular topic in language testing research as well as in second language learning. Second language learners start to learn language differently from the way first language learners do. Second language learners already have linguistic resources of the first language to express their ideas while first language learners do not. The knowledge of first language can be a basis to learn second language. That is, transfer may occur. This can either accelerate or hinder the second language learning (Zobl, 1982). Since some languages are structurally closer to the target language than others, the learners whose native languages have similar structure to the target language would learn the target language faster than the learners whose

language structure is different from that of the target language. If this is true, second language mastery may appear differently across different native language groups.

Much research has been done from various perspectives to show the effects of native language on second language proficiency tests. Factor analyses (Dunbar, 1982; Sawatdirakpong, 1993; Swinton & Powers, 1980) have been used to study the relationship between test takers' native languages and internal structure of the performance on language proficiency tests. Item-bias analyses (Ryan and Bachman, 1992; Alderman & Holland, 1981) have been used to identify what cause advantages for or disadvantages against one group by studying group performance on an item. Other researchers (Spurling & Ilyin, 1985) grouped their subjects according to test taker characteristics such as nationality, learning styles, major field, gender. They then compared overall test performance between different groups.

Swinton and Powers (1980) conducted a factor analysis of the TOEFL for seven language groups; African, Arabic, Chinese, Farsi, German, Japanese, and Spanish. This study was to provide evidence of construct validity of the TOEFL by determining precisely what component abilities the test measures i.e., the explanatory constructs that account for examinee performance. They found that a four factor solution (one general factor and three secondary factors) was appropriate for explaining most of the variability of examinees' performance. For the listening comprehension

section, the great majority of items loaded on a single factor in each language group. The result of the structure, written expression, vocabulary and reading comprehension section appeared somewhat more complex. For Chinese, African, Arabic, and Japanese groups, the structure, the written expression, and the reading comprehension section loaded on one factor, and the vocabulary on a different factor. For German and Spanish groups, the structure and the written expression loaded on one factor, and the vocabulary and the reading comprehension were loaded on a different factor.

They also found that the vocabulary factor was most likely to form a separate factor from the listening comprehension factor. The vocabulary factor was positively correlated with age and degree-intention of examinees in every language group. This implied that the vocabulary scores may result from training and experience, and that it may have to be reported separately.

Swinton and Powers recognized the implications of their study for the interpretation of TOEFL subscores. For the non-Indo-European group, the scores on structure and written expression did not match the scores on vocabulary and reading comprehension. They inferred that this may have resulted from the lack of knowledge of English vocabulary rather than lack of reading comprehension ability compared to the Indo-European group. They inferred also that low scores on the vocabulary and reading comprehension may not have been critical for non-Indo-European group since vocabulary could

have been learned more readily than grammatical or syntactical structure.

In his confirmatory factor analysis of the internal structure of the TOEFL for seven native language groups, Dunbar (1982) also found that a four factor model (a general factor and one factor corresponding to each of the three TOEFL sections) fitted the data best. In this study, the general factor appeared dominantly in all groups but intercorrelation between factors appeared somewhat different between groups. For the African group, intercorrelations among factors II, III, and IV were high. For the Arabic, Chinese, and Germanic groups, factor III correlated moderately with factor II and IV. This showed that first language played a different role for each language group.

Oltman et al. (1988) studied the influence of examinees' native language in relation to their language proficiency on the TOEFL. They used three approaches to multidimensional scaling to study the interrelation among TOEFL items, varying with native language and language proficiency. They identified four dimensions. Three of them corresponded to each of the TOEFL's three sections. These dimensions consisted mainly of the easy items of each section. The fourth was associated with the difficult items in the reading comprehension section, and appeared to be an end-of-test phenomenon. The three dimensions of easy items appeared more salient for the low-scoring subsamples, but did not differ across the language groups. The end-of-test phenomenon

appeared more salient for some language groups. "The similarity in the dimensions for the different language groups suggests that the test is measuring the same construct in each group" (p. 29). They also recognized the further research needs to investigate the cause of the language group difference in the end-of-test dimension.

Other research has tried to account for the source of differences in learners' performance on individual items which appear easier for certain groups. Kunnan (1990) conducted a DIF (differential item functioning) study on the English as a Second Language Placement Test takers at the University of California at Los Angeles. He used the one-parameter Rasch model to compare group performance on items according to their native language and gender. Among the many language groups, four large ones were compared: Chinese (262), Spanish (81), Korean (76), and Japanese (59). The subjects were also grouped into the male group of 478 and the female group of 347. The male group was 478 and the female group was 347.

He found 13 DIF items in the native language group analysis, and 23 DIF items in the gender group analysis. He inferred that the possible sources of the difference might be due to the differences in the instructional backgrounds of the subjects as well as linguistic affinity. All the items which appeared favorable for the Spanish group were vocabulary problems such as 'hypothetical', 'implication', 'elaborate', and 'alcoholics'. All these words were cognates

that Spanish shares with English, which made the items easier for the Spanish group. He also found that many items functioned differentially across genders. He suspected the difference might have resulted from the content. Male-favored items were from the passages from business, culture/anthropology, and engineering in which males seemed to outnumber females.

A similar study was done by Ryan and Bachman (1992). They conducted a DIF study on the performance on the two most widely used tests, TOEFL and FCE (First Certificate of English). The subjects took the both tests. They grouped the subjects according to their first language (Indo-European (n=792) and non Indo-European (n=632)) and gender (a male group of 575, and a female group of 851). The result of the DIF study was compared to *a priori* judgment of content rating. Gender difference did not result in mean differences on both tests. On the TOEFL, the Indo-European group performed considerably higher. On the FCE vocabulary test, neither language group showed a difference. However on the FCE reading test, the Indo-European group gained a higher mean. Content analyses showed that the non-Indo-European group favored items in the TOEFL which tended to be "more specific in terms of their American cultural, academic and technical content than the items which favored Indo-European native speakers or which showed no DIF" (p.22). They suggested this phenomenon may be related to the differences in their test preparation.

In their study, Alderman and Holland (1981) compared test performance and item performance across several language groups on two administrations (November 1996 and November 1979) of the TOEFL: Japanese, African, Arabic, Chinese, German, and Spanish. They took about 1000 examinees from each language group. They found that Spanish and German groups did better in all three sections of the both administrations than other groups, and that test takers of comparable scores from different groups differed in their performance on specific items according to their native language. They then tested whether the differences in performance between groups could be explained by linguistic similarities and differences. They asked ESL specialists to review the result of the first administration and then to identify probable items of discrepant performance across groups. However, a priori prediction based on linguistic contrast turned out to be unreliable. They noted that native language surely had an influence on acquisition and performance in a second language, but that it is not clear how much language proficiency tests reflect linguistic affinity.

Extensive research on the effects of learners' variables was done by Spurling and Ilyin (1985). They studied how the learner variables of age, sex, language background, high school graduation and length of stay in the United States affected the learners' performance on six tests: two cloze tests, a reading test, a structure test, and two listening

tests. Their subjects were 257 students enrolled at Alemany Community College Center, an adult education center in San Francisco. They found that the variables of high school graduation status, first language, and age affected learner's performance significantly. The other three variables did not appear significant. They also found that certain tests appeared favorable to certain groups. They gave a warning that one should be careful in interpreting test results and interpretation should be based on the objective of the test. If the objective was to test particular skills, a set of tests could be considered independently. If the test was to measure overall language proficiency, however, simply adding the results of subtests could be biased for or against certain groups. They also suggested that a weighting of the subtests be considered in the decision making process.

In sum, studies showed native language plays an important role in constructing learner's language proficiency. However, it was not clear to what degree native language affects test performance on proficiency tests. The internal factor structures of test performance appear different across different language groups (Dunbar, 1982; Swinton & Powers, 1980). Results has shown that language test performance can be explained more than one factor. Swinton and Powers (1980) and Dunbar (1982) showed that factor loadings varied across different language groups, that is, the internal structures of language test performance of different language groups were different. Oltman et al.

(1988) speculated that the difference between their results and the other research may be because other studies did not consider language proficiency. They found that test performance of higher level students did not display equally performance on three easy item dimensions as that of lower level students.

Item performance (Alderman & Holland, 1981; Ryan and Bachman, 1992) and overall mastery profiles (Alderman & Holland, 1981; Ryan & Bachman, 1992; Spurling and Ilyin, 1985) appeared different across different language groups. The studies of the TOEFL (Ryan & Bachman, 1992; Alderman & Holland, 1980) showed that the Indo-European group performed better than other language groups in all three sections, but in their study on the FCE, Ryan and Bachman (1992) found no difference in the vocabulary section between the Indo-European group and the non-Indo-European group.

2.2.2 Native country and language instruction

Another popular test taker characteristic in language testing research has been native country. Native language and native country do not necessarily correspond. Some countries such as Switzerland and Canada have more than one official language. Some languages such as English and Spanish are spoken in more than one country.

Each country has its own educational policy depending on its educational culture and needs. That is, language instruction in some countries may follow different instructional approaches and emphasize one language skill

more than others (Farhady, 1982). Students may naturally have different mastery profiles of second language skills across different countries.

It seems obvious that test takers' mastery profiles are dependent more on type of learning experience than nationality. However, little research has been done to investigate directly the effects of instructional methods on constructing language proficiency, even though a few researchers (Carroll, 1961; Farhady, 1982) have reasoned that nationality effects may be due to differences in instructional methods.

Hosley (1978) studied the effect of country of origin and sex on the TOEFL. The researcher examined the 147 subjects who enrolled in the Center for English as a Second Language (CESL). The subjects were from Mexico, Saudi Arabia, Libya, Venezuela, Japan and others. In this study, the subjects from Mexico performed best and the subjects from Saudi Arabia and Libya worst. The researcher also identified the source of most difference from the listening comprehension and the vocabulary sections. However, the effect of sex did not appear significant. From this study, he assumed that learning experience they had in their home countries may have resulted in such differences.

The effect of learning experience was also implied in Politzer and McGroarty's study (1985). They examined the relationship of learning styles to gains in English language learning. They grouped 37 students who enrolled in an eight

week intensive English course according to their cultural backgrounds (Hispanic vs. Asian) and field of specialization (professional engineering and science vs. social science and humanities). They gathered the information of classroom behavior, individual study, and interaction based on the questionnaire. They then identified desirable learning styles by matching the questionnaire information and gain scores of the learners on four tests. They found out the Hispanic group had more desirable learning styles[2] (The Hispanic group scored higher than the Asian group on the questionnaire scale. This could be because most of questionnaire items were related to social interactions such as correcting fellow students, asking teachers all kind of questions, and asking for help and confirmation). The Asian group, even though they had less desirable styles, achieved more gains than Hispanic group in linguistic competence and communicative competence tests (the Comprehensive English Language Test for Speakers of English as a Second Language: Harris and Palmer, 1970) whereas Hispanic groups achieved more in oral proficiency test and auditory comprehension test (The Plaister Aural Comprehension Test: Plaister & Blatchford 1971). However, the comparison by

[2] They obtained the information about the learning style from the questionnaire they designed on the basis of their survey of the available literature on behaviors and strategies of good language learner. However, as they stated in their study, these characteristics of good language learning behavior was not based on a unified theoretical perspective. They treated these characteristics as heuristic constructs, and calculated internal consistency of students' responses with Cronbach's alpha coefficient after scaling the responses. They then eliminated 19 items which showed negative bisirial correlation with the total scale. High scores on the scale were equated to good learning styles.

academic field overlapped with the comparison by cultural background. Most of the engineering and science students were Asian and all of the social science and humanities students were Hispanic. Different learning styles between Asians and Hispanics may reflect the type of previous English instruction the subjects had received. Many of the Asian countries emphasized "rote memorization, translation of texts, or recognition of correct grammatical forms in reading" (p.114).

There has been little research which compared the efficiency of instructional methods or learning experience. One of the studies was done by Landolfi (1991). Landolfi investigated whether or not a methodological change had a real measurable effect on achievement in educational tests. She studied two school districts in Los Angeles which changed their bilingual program, shifting from a grammar-based syllabus to a comprehension-oriented one. That is, the focus of the bilingual program shifted from structures to comprehension.

The grammar-based approach presented the language as a puzzle and taught one piece at a time. Only the teacher knew how the whole picture looked like until all grammatical point were taught to students. The students developed metalinguistic knowledge by studying about the language. On the other hand, the comprehension-based approach presented the language as a whole picture and the students broke down the picture into small pieces.

The achievement scores in the CTBS (Comprehensive Tests of Basic Skills) were compared under two types of instruction, before and after the change. The CTBS contained a series of batteries of tests (reading, language, social studies, science and mathematics) from kindergarten through grade 12. The outcomes of the first two components, reading and language were analyzed. The tests were grammar-oriented test, basically designed for native speakers. Their focus was on English phonology, syntax, semantics and rhetoric. Data from 480 students from grade 1 to grade 3 of both groups were gathered.

In the comparison of the data of the first and second grade students, grammar students did better in both the reading and language components. However, the differences disappeared by the end of year three. That is, by the third grade, the comprehension-trained students achieved the same results as the grammar-trained students without being exposed to explicit training of grammar learning.

Both groups in this study learned language in a natural setting as well as at school. This may have made it possible for both groups of students to attain the same level on the CTBS partly because their mastery of language may have been at the same level regardless of the learning experience and partly because the proportion of comprehension-type questions increased as the grade changed. The result would not be the same, however, if the situation had been an EFL setting. A study on the effects of such different settings follows.

Mitchell (1991) paid attention to the sociolinguistic contexts where students had learned English. She divided learning environments into two types, ESL (English as a second language) and EFL (English as a foreign language) environments. In an ESL context, English was one of several languages used on a daily basis. Students learned English in a natural environment as well as in a school. In an EFL context, learning English was limited only to a classroom. Students from an EFL context might have different patterns of strong skills and weak skills from ESL students. They may also need different preparation for studying in America. However, these differences were not considered when international students were assigned to ESL courses. They were grouped together according to their scores regardless of the learning contexts where students had learned English.

Mitchell examined whether students' performance on the 3 subsections (structure, cloze, and dictation) of the UIUC EPT varies according to the contexts where students had learned English. Total of 146 subjects representing 25 different countries were analyzed. Half of them were from EFL contexts and the other half were from ESL contexts.

Two analyses were done. The first analysis dealt with all students of all countries. In the second analysis, only eight students of each of the three most represented countries of each context were chosen. In this study, she found that there was a significant interaction between environment and test types. EFL students performed better in

the structure section and ESL students better in dictation. There was not a significant difference in the cloze section. Similar results were found in the second analysis. She explained that this was because EFL students had had more experience with books and exams on grammar and that ESL students had developed the capacity to function communicatively in English. This suggests that the placement instruments may need to be examined in order to more accurately match the specific needs of students from ESL contexts with their course placements[3].

Insensitivity to students' learning experience was pointed out by Farhady (1982), as well. He argued that the test takers of language proficiency tests like the UCLA ESLPE (English as a Second Language Placement Examination) were not homogeneous and that the definition of a proficiency test should have included test taker's characteristics as potential dimensions in language testing. Learners were not homogeneous in their proficiency. Learners from different educational backgrounds had certain performance profiles which indicate strengths and weakness in different language skills.

To show heterogeneity of learners' dimensions, Farhady studied how learner variables affected performance. He took the 800 students' scores on the UCLA ESLPE. The UCLA ESLPE consisted of five sections: cloze, dictation, listening

[3] The cloze and dictation have been dropped and a video-essay test added, one which is more relevant to matching students' needs with course placements. Mitchell's work was a key feature in motivating that change

comprehension, reading comprehension and grammar. He grouped the students according to sex, university status, nationality, and major field of study. He looked into how the groups performed in different sections of the test. He found that those five variables were significant factors which accounted for the performance differences between groups in one or more sections of the ESLPE.

This study showed that learners had different degrees of mastery in different sections according to the test taker characteristics. Most existing ESL tests like the UCLA ESLPE did not take into account these variables. This may have led the test to fail to assess learners' needs accurately. As a result, opportunities of more efficient instruction may have been lost.

2.3. Implications for Research

As a process of construct validation, understanding student's backgrounds is very important. It provides a basis for using and making inferences from test results. The main purpose of giving an EPT (ESL Placement Test) is to diagnose students' weaknesses in English proficiency and place them in appropriate ESL courses, which are designed to prepare students for their study in English medium universities. However, assessing students' needs simply based on the EPT results may not be appropriate. Performance on an EPT is not independent of the ways in which language is acquired. Learners are not homogeneous. They have various backgrounds; different nationality, different native languages, different

learning styles, different culture, different learning experience etc. Their various backgrounds may have affected their English language learning. With the understanding of the effects, one can provide a better account of differences across various groups and a better diagnosis of the nature of the problems learners might have, thereby providing more efficient instruction.

Studies have demonstrated the importance of considering students' various background factors in interpreting test results. Those studies provided evidence of effects of various characteristics on test performance. However, little research has been done to study the relative effects between test taker characteristics, that is, in what degree language test performance is affected by one characteristic in relation to another. The present study investigated the effects of two characteristics, native language and learning experience, then assessed their relative effects. These two characteristics were chosen for two reasons. First, these two characteristics have been shown the most influential on test performance from other studies. The second reason is a practical concern. It may not be helpful to consider all the characteristics in using EPT test results. ESL programs usually do not enjoy enough budget and the number of students grouped by a certain variable is usually not enough to make a class. Including one or two characteristics into the EPT administration may be enough for most situations.

Previous findings on the effect of native language showed that the internal structures of language test performance are different across different language groups and that Indo-European language group generally tend to have a higher proficiency profile than other groups. In the TOEFL studies of Ryan and Bachman (1992) and Alderman and Holland (1980), Indo-European language groups performed best in all sections. Spurling and Ilyin (1985) found that the Spanish group performed better in their cloze, reading, and structure test than Chinese or Vietnamese group. In the studies of test performance on the FCE, Ryan and Bachman (1992) found that Indo-European group performed better on the reading section than non Indo-European group, but that two groups did not perform differently on the vocabulary test.

While studies of first language almost invariably found superior performance of Indo-European language group, effects of learning experience were mixed. Landolfi (1991) found that though a grammar-based syllabus was more efficient in grammar teaching than a communication-based syllabus at the beginning, the two syllabi did not show differences after three years of teaching. Mitchell (1991) found that the learning environment was also important. In her study, EFL students performed better in the structure section while ESL students performed better in dictation. Farhady (1982) also found significant interactions between native country and section scores on UCLA ESLPE. Politzer and McGroarty (1985) identified the learning styles of Hispanics and Asians and

found that the Hispanic group achieved more in oral proficiency tests and auditory comprehension tests whereas their Asian group performed better in linguistic and communicative competence tests.

These findings imply that native language seems to have more influence on the performance on language proficiency tests than learning experience. When different native language groups were compared, the Indo-European language group (to which English belongs) performed better in almost all types of test. When test takers were grouped by learning experience, significant interactions were found in most studies. However, this does not decide which learner characteristic is more influential. What hasn't been known is which characteristic is more influential when only scores from one test section scores are analyzed. Even though native language seem to have a uniform effect on overall tests, learning experience might show bigger effects when the effects of the two characteristics on only one section are compared. If this is true, scores of the section should be interpreted and used with consideration of learning experience. If native language turns out more influential on each section as well as overall tests, learner's native language should be given more weight in using the test results. As an effort to understand and use the test performance correctly, this study will try to find answers to a question: Which characteristic should be given more weight in interpreting language test performance on the UIUC EPT?

CHAPTER III. BACKGROUND INFORMATION AND RESEARCH QUESTION

3. 1. UIUC EPT

The EPT (English as a Second Language Placement Test) of University of Illinois of Urbana-Champaign is given to new international students whose TOEFL scores are below a certain campus or department requirement (UIUC requires the EPT for students below 607, unless a department has a higher cutoff value). It is designed to test whether test takers have an appropriate level of academic English proficiency for them to study at UIUC and to diagnose the problems they might have. The UIUC EPT consists of three sections: structure, video-essay and pronunciation.

The structure section[4] is designed to test the knowledge of English grammar and expressions. It consists of 50 multiple-choice items. Each item has one correct answers and three distractors. No penalty is applied for guessing.

The video-essay section is designed to test the ability to integrate information from two modalities and use it in an essay. Students are given a video taped lecture followed by a passage to read on the same topic and asked to write a short essay based on the lecture and passage. The content and the format of the video tape is a simulated part of a ordinary classroom lecture. Students may take notes as they do in the actual classroom. Essays are graded from level one to four; Level one is the lowest and level four the highest. On the

[4] The entire structure section is not allowed to be reprinted in this thesis for security reasons. A few sample items are provided in the discussion section.

basis of the structure section scores and grades from the video-essay section, students are assigned to various level of ESL classes.

The pronunciation section is designed to test whether students can communicate intelligibly in the classroom. Students are interviewed and asked to read dialogues, paragraphs, and sentences. Some of the sentences have difficult words which are probably new to the students. Students' pronunciation of each syllable, stress, intonation, and latent ability to put stress on a new word are checked. Their ability is judged as 'Required,' 'Recommended,' and 'Pass.'[5] 'Required' means the student has to take ESL 110 which is designed to improve the pronunciation. Students who get 'Recommended' do not have to take ESL 110, but are recommended to take it. 'Pass' means that the pronunciation is very intelligible and the students do not need to take ESL 110.

3.2. Native Language

A topic of this study is the effect of the linguistic affinity of a language to English on the performance on the UIUC EPT. Since a language group is usually represented by a small number of test takers, it is reasonable to group languages according to their linguistic affinity. The classification of native language of this study followed the

[5] Original grades given by interviewer are from 1, the lowest, to 5 the highest. These grades are converted to three categories: '1' or '2' = 'Required', '3' = 'Recommended', and '4' or '5' = 'Pass. 'This study did not use the original grades because of the small number of subjects

genetic typology of language family (Grimes, 1992). Languages in a language family share cognates, similar phonological and syntactical structure.

The UIUC EPT administration acknowledges 61 native languages of students. Fifty five languages represent almost all EPT takers. Those languages can be classified into 10 language family groups; Indo-European, Afro-Asiatic, Austro-Asiatic, Sino-Tibetan, Uralic, Niger-Congo, Austronesian, Dravidian, Altaic, and Daic. Table 3.1 is the classification of language families.

Table 3.1. Classification of language families

Language Family	Languages
Altaic	Japanese, Korean, Turkish
Sino-Tibetan	Cantonese, Mandarin
Indo-European	Armenian, Assamese, Awadhi, Bengali, Bhojpuri, Bulgarian, Danish, Dutch, English, French, German, Greek, Gujerati, Hindi, Italian, Marathi, Nepali, Norwegian, Oriya, Persian, Polish, Portuguese, Punjabi, Pushto, Romanian, Russian, Saraiki, Serbo-Croatian, Spanish, Swedish
Afro-Asiatic	Amharic, Arabic, Hebrew
Austro-Asiatic	Cambodian, Vietnamese
Niger-Congo	Igbo, Nigeria, Lozi, Swahili, Yoruba
Austronesian	Indonesian, Maly, Tagalog
Daic	Tai, Lao
Uralic	Finnish, Hungarian
Dravidian	Kannada, Malayalam, Tamil

3.3. Language Learning Experience

The most important factor which determines one's education is probably one's nationality. A learner's language learning experience is mainly dependent on his/her native

country's own situation: educational objectives, economic ability, and social needs. Grouping by nationality, however, does not fit to the interest of this study. It would disregard the similarity of instructional methods between countries and diversity of learning experience within a country. As an alternative, this study investigated the three most widely used instructional methods; (1) grammar-and-reading-focused instructional method, (2) controlled-oral-language-focused instructional method, (3) communication-skill-focused instructional method. Each has a different view on language learning and emphasize different language skills.

The grammar-and-reading-focused instructional method dates from the Renaissance when Latin and Greek literature was taught (Celce-Murcia, 1991). The main interest of this method is reading and interpreting the meaning of texts. Spoken language is not regarded as important. Most of classroom activities consist of reading and translating texts. Teachers do not have to have special skills to teach and they can handle large-sized classrooms easily (Brown, 1987). They do not have to be fluent in the spoken language.

As an anti-grammar-and-reading-focused method, the controlled-oral-language focused instructional method is based on behavioral psychology and structural linguistics. Two main principles of the method are (1) "language is speech not writing," and (2) "language is a set of habits" (Diller 1971). Thus teachers focus little on written language skills. Most classroom activities are memorization and automatization

of the expressions of the target language. Students are not allowed to produce unlearned expressions. They always have to imitate exactly what they have heard.

While the controlled-oral-language focused instructional method emphasizes spoken language skills such as speaking and listening, the communication-focused instructional method emphasizes both written and spoken language skills, reading, writing, listening, and speaking. It views language as a means of communication and language learning as a process of internalizing target language rules. Compared to the other two methods, it requires the most tolerance about errors from teachers because making errors is regarded as a part of the language learning process.

This categorization may not reflect all existing instructional methods. Instructional methods of some test takers may not be categorized definitely into one of these types. They may have features of two or all of the above methods, or others.

3.4. Research Questions

This study examined and compared the extent to which two test taker characteristics, native language and learning experience, affected test performance on the UIUC EPT. Learning experience was operationalized by surveying the test takers. Two levels of analyses were conducted. First, overall proficiency profiles across groups of different language groups were compared. Second, item level analyses were conducted to examine what type of items function

differentially across groups. The item level analyses were limited to the structure section because it is the only multiple item part of the EPT, for which data are available.

This study addressed five research questions:

Research question A: Are language proficiency profiles different across different language groups? Since English belongs to the Indo-European language family, which means that languages of Indo-European family share similar structures and cognates, the Indo-European group is expected to do best of various language groups, at least at the structure section.

Research question B: Are language proficiency profiles different across different learning experience groups? Different instructional methods emphasize different skills. The learning experience of a group may be focused more on one skill than on the others. The grammar-and-reading-focused learning experience group is expected to perform best on the structure section. The controlled-oral-language-focused group and the communication-focused group are expected to perform better at the pronunciation section than the grammar focused group. The communication-focused group is expected to perform better on the writing section than the other two groups.

Research question C: What types of items on the structure test function differentially across different language groups or are biased against one language group[6]*?*

[6] Whether an item or a group of items is biased or functioning differentially depends on the validity of the items. When the items are judged as not valid, that is, the items are measuring something other than the target knowledge, the items are said to be biased.

Since analyses of differential item functioning was conducted only with the structure section, more items were expected to appear favorable to Indo-European groups than items which favor any other groups. Those items were expected to reflect the similarities and differences between different languages.

Research question D: What types of items on the structure test function differentially across different learning experience groups or are biased against one learning experience group? The grammar-focused group was expected to have more favorable items than any other groups.

Research question E : Which of the two characteristics can explain test takers' performance on the UIUC EPT better? The answer to this question will imply which variable is more crucial in interpreting and using the test results.

Otherwise, the items are said to function differentially across different groups. More discussion is in 4.3.

CHAPTER IV. METHOD

4.1. Subjects

Among the newly admitted students to UIUC for the fall semester of 1994, 315 students were asked to take the EPT. The EPT takers were asked to fill out a questionnaire[7], which was constructed to obtain information about students' native languages and learning experience. Two hundred fifty five students returned the questionnaire.

The range of English proficiency of the subjects was assumed to be limited because the EPT takers had to submit their TOEFL score to prove that they had over a certain level of proficiency before they got admitted, and because students with over a certain TOEFL score were exempted from taking the EPT.

4.2. Grouping

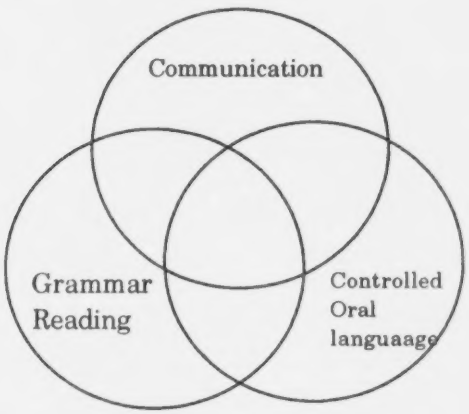
The subjects were grouped according to their first language and learning experience. The information on students' characteristics was obtained from the questionnaire. The questionnaire consisted of two parts. The first part asked native language, native country, major field of study, and academic status. The second part which investigated learning experience consisted of 11 questions. Each asked whether test takers had experienced distinctive features of one or two instructional methods. Because the students could have experienced more than one instructional

[7] Appendix A.

method, it was explicitly stated that the questions were about the information of the institutes where the learner studied most of his or her English.

It is very unlikely that a learner has learned English only under one instructional method. Probably, the learners have experienced every feature of all three instructional methods in different degrees. What should be determined to categorize students' learning experience is which instructional method has been dominant. This is easily understood in a diagram. In figure 4.1. three types of learning experience are represented by three circles. An examinee's learning experience can be represented as a point within the diagram. If the students have experienced one major method, his experience will be placed on any of three non-overlapped area. Likewise, if he has experienced two or more major methods, he will be placed on an overlapped area.

Figure 4.1. Classification of learning experience. This model represents students' dominant learning experience



The analysis of the questionnaire was done by adding up the features the examinees had answered that they experienced, and comparing the number of features for each instructional method. The total number of features was 11. The method of the highest number was taken as the major instructional method for each examinee. For example, if a student answered that s/he experienced 3 features for grammar and reading focused method, 5 features for the controlled oral language focused method, and 4 features for communication focused method, his or her dominant experience was categorized as the controlled oral language focused method. Table 4.1. is the summary of the result of all 255 examinees' learning experience judgments. Fifty two Examinees were judged to have experienced the same number of the features of two or all three instructional methods. That is, these examinees belonged to the overlapped area in the figure 4.1. They were excluded from the remaining analyses. Eighty one examinees were judged to have had the grammar-and-

Table 4.1. The summary of the result of examinees' learning experience judgment

Learning Experience	*f*	%	cum *f*	cum %
Grammar and Reading	81	31.8	81	31.8
Controlled Oral Language	76	29.8	157	61.6
Communication	46	18.0	203	79.6
Grammar & Controlled*	10	3.9	213	83.5
Grammar & Communication*	2	.8	215	84.3
Controlled & Communication*	37	14.5	252	98.8
All three methods*	3	1.2	255	100.0

'*' 52 examinees experienced the same number of the features of two or all three instructional methods.

reading-focused method, 76 examinees were judged to have had the controlled oral language-focused method, and 46 examinees answered they had had the communication-focused method.

Students were also grouped according to their first language[8]. Since some language groups were represented by small number of examinees, language family was used to group the examinees. This resulted in 3 major language family groups and 5 minor language family groups. Among the total 255 subjects, 224 subjects belonged to 3 major groups; Altaic, Sino-Tibetan, and Indo-European. Thirty one were grouped into one of the minor language families; Afro-Asiatic, Austro-Asiatic, Niger-Congo, Austonesian, and Daic[9]. These minor groups were excluded from the analyses due to the small numbers of subjects. Table 4.2. is the summary of language family groups.

Table 4.2. Language families and number of subjects

Language Family	Number of students
Altaic	70
Sino-Tibetan	72
Indo-European	82
Afro-Asiatic	5
Austro-Asiatic	1
Niger-Congo	1
Austronesian	15
Daic	9

[8] The present study was interested only in first language. It did not consider how importantly and extensively English had been used as a second or third language for the subjects.

[9] See Table 3.1 for the classification of language families.

After small groups were removed from both the learning experience grouping and first language grouping, subjects were matched with the UIUC EPT data set to obtain scores of each subjects, yielding 203 subjects for the analyses reported in this study. They were distributed as in Table 4.3.

Table 4.3. Number of examinees by language families and instructional methods

	Grammar & Reading	Controlled Oral Language	Communication	Total
Indo-European	13	16	41	70
Sino-Tibetan	28	15	24	67
Altaic	32	14	20	66
Total	73	45	85	203

4.3. Analyses

To compare group performance in each section, an ANOVA and a Chi-square statistic were used according to the nature of scores. A two-way ANOVA was used to study the effect of native language, learning experience and interaction of both variables on the structure section. On the pronunciation section and the video-essay section, the subjects were graded categorically; 'Required', 'Recommended,' or 'Pass' for the pronunciation skill and grade 1 to grade 4 for the writing skill. A Chi-square statistic was used to compare group performance on the pronunciation section and the video-essay section.

For the item level analyses, the SIB (simultaneous item bias) test was used to study how groups performed differently

for an item or a group of the 0/1 scored items from the EPT structure test. The SIBTEST is based on a multidimensional item response theory model of test bias (Stout and Roussos, 1992, p. 1). It detects unidirectional DIF (Differential Item Functioning) or bias (Stout, & Shealy, 1992, p.14) when an item or a group of items favor all members of one group over another group. When bias or DIF effects are crossed, for example, when an item appears favorable for higher level examinees of one group and favorable for lower level examinees of the other group, the SIBTEST can only detect the amount of bias or DIF beyond the amount of cancellation.

There was some concern about the sample size. Because SIBTEST is based on asymptotic distributions, a large sample size per group was required. However, the subjects sizes of each group in the present study ranged from 45 to 85. Due to the small sample sizes, Type I error would have been higher than .05 which was the criterion used for flagging DIF items in the present study.[10]

The SIBTEST produced three types of output: (a) individual DIF which was the magnitude of DIF of each item, (b) group DIF which was the collective amount of DIF of a group of items, (c) DTF (Differential Test Functioning) which was the collective amount of DIF of all items. The SIBTEST produced DTF by calculating a cancellation effect among DIF

[10] A minimum sample size is still under study. Ackerman said that at least 150 subjects per group would be required (personal communication, October, 1994). In their simulation study, Roussos and Stout (1996) showed that with small sample sizes of 100, 200, 500, and 1000, the SIBTEST maintained the nominal level of significance (.05).

items (Stout and Roussos, 1992). Calculating group DIF and DTF is a unique advantage of the SIBTEST over other tests. Because SIBTEST can compare only two groups at a time, 6 comparisons were made; 3 pairs of different first language groups and 3 pairs of different learning experience groups.

Before conducting item level analyses, the DIMTEST program (Stout, 1987; Stout, Douglas, Junker, & Roussos, 1993) was run to see if the data are essentially unidimensional (Stout, 1987; Nandakumar, 1991), i.e., if there is only one dominant dimensions in the data. If there is one dominant dimension with several minor dimensions in the data, it can be said that the test are measuring one ability. Items related to minor dimensions can be either bias items or DIF items depending on the validation of the items[11]. If more than one dominant dimensions are found and the validity of items is in question, the items are said to be biased. If more than one dominant dimensions are found and the items related to each dimension are valid, the target ability can be said to be multi-dimensional.

---

11 Two terms, bias and DIF, need to be clarified in relation to dimensionality. Bias means that an item or a group of items is less valid for one group of examinees than another group of examinees on an intended target ability. Performance on these items is affected by knowledge other than that of the target knowledge. These items constitute a secondary or tertiary dimension. DIF is the notion that an item or a group of items favor one group of examinees over another group of examinees without referring to the concept of validity. These items could also constitute a secondary or tertiary dimension, but the items or the dimensions are not disvalidated or validity is not in question.

CHAPTER V. RESULTS

5.1. Group Differences in Each Section.

5.1.1. Structure

Before conducting the ANOVA, the reliability of the structure section scores was obtained and three assumptions of ANOVA were checked; independence of observations, normal distribution of the dependent variable, and equal variance.

The reliability was measured using Cronbach alpha. The reliability of the structure section was .74, which was low, compared to other standardized tests like the TOEFL (F. Davidson, personal communication, June, 1996). This may have been partly due to the limited range of English proficiency of the subjects.

Among the three assumptions, independence of observations was not in doubt because the subjects took the test independently. Two other assumptions were checked. Table 5.1. is the summary of descriptive statistics of the structure scores for each language and learning experience group. Shapiro-Wilk statistics in Table 5.1. indicated that the normality assumption for all groups except Altaic and grammar-and-reading-focused learning experience group were tenable at the significance level of .05. However, ANOVA is known to be robust to violation of the normality assumption (Kirk, 1995).

Cochran's *C* test was used to test homogeneity of population variances of all groups. Cochran's *C* test did not

reject homogeneity of population variances (C = .1758, df=9, 15, $p<.01$).

Table 5.1. Summary descriptive statistics of structure scores by language family and learning experience.

Language	Experience	n	Mean	Variance	W:normal	P < W
IE	GR	13	50.23	62.69	.96	.79
IE	CO	16	43.93	114.86	.93	.28
IE	COMM	41	49.12	92.29	.95	.10
ST	GR	28	48.07	67.84	.96	.47
ST	CO	15	48.13	72.83	.92	.19
ST	COMM	24	49.45	59.82	.93	.16
Altaic	GR	32	52.13	46.62	.91	.01
Altaic	CO	14	49.43	76.10	.94	.44
Altaic	COMM	20	50.65	60.23	.97	.78

Table 5.2. Descriptive statistics by language family.

Language	n	Mean	Variance
IE	70	48.14	94.82
ST	67	48.58	64.49
Altaic	66	51.10	56.25

Table 5.3. Descriptive statistics by learning experience group.

Experience	n	Mean	Variance
GR	73	50.23	59.37
CO	45	47.04	90.54
COMM	85	49.57	74.29

Table 5.2 and Table 5.3 are the descriptive summaries by language family and language learning experience group, respectively. Glancing at the group means, the Altaic group among the language family groups and grammar-and-reading-focused group (GR) achieved the highest scores. However, group differences were relatively small when the collective amount of variances are considered. This was confirmed by the

ANOVA. Table 5.4 is the ANOVA summary table. No difference was detected at the significance level of .05. That is, performance on the structure section did not differ according to examinees' group memberships. The effect sizes for native language, learning experience and interaction of the two were less than .01[12] ($f_{\text{native language}} = .094$; $f_{\text{learning experience}} = .090$; $f_{\text{interaction}} = 0$).

Table 5.4. ANOVA result of the structure section.

Dependent Variable : scores on the structure section.

Source	DF	SS	MS	F	P > F
Language	2	273.53148	136.76574	1.90	0.1525
Experience	2	263.46132	132.73066	1.83	0.1633
Language * Experience	4	212.94428	53.23067	.74	0.5633
Error	194	13970.66282	72.01373		
Total	202	14798.18719			

5.1.2. Pronunciation

A chi-square statistic showed that there were significant differences between groups in the performance on the pronunciation section. However, differences were not detected among the different learning experience groups. Table 5.5. is the result of the chi-square analysis when the examinees were grouped by their language family and by their learning experience. Table 5.6. is a summary of cell Chi-square analysis for three language groups. The Indo-European

12 Cohen provided the guidelines for interpreting the *f* measure of effect size:
f = .10 is a small effect size
f = .25 is a medium effect size
f = .40 or larger is a large effect size (as cited in Kirk, 1995, p. 181)

(IE) group appeared to have performed best of three groups and the Sino-Tibetan (ST) group performed worst. Eighty four percent of IE examinees got a 'Pass' degree and only 1.43 % of the IE examinees were required to take the ESL pronunciation class. 40.3 % of the ST examinees (which was the largest proportion of people among three corresponding groups) were required to take the pronunciation class and 32.84% of the ST examinees passed the test. The Altaic group stood between the IE group and the ST group. The cell chi-square statistics also implied the same relative standings of three groups. They indicated where the differences were. The biggest contribution (20.66) to the chi-square was found among the ST examinees who got 'Required.' That is, more examinees of the Sino-Tibetan group got 'Required' than the expected frequency and fewer examinees got 'Pass' than the expected frequency. The IE group appeared in a reverse way. More examinees got 'Pass' than the expected frequency and less examinees were required to take ESL 110 than the expected number of examinees. Cell Chi-squares of the Altaic group were relatively small compared to other groups.

Table 5.5. Chi-square analysis on the pronunciation section

Grouping	DF	Chi-square	Prob
Language Family	4	50.472	.000
Learning Experience	4	2.729	.604

Table 5.6. Cell chi-square analysis for the pronunciation section when the examinees were grouped by language family.

Grade	Statistics	IE	ST	Altaic	Total
Pass	Frequency	59	22	46	127
	Expected	43.793	41.916	41.291	
	Deviation	15.207	-19.92	4.7094	
	Cell Chi-Square	5.2805	9.4631	0.5371	
	Column percent	84.29	32.84	69.70	
Recommended	Frequency	10	18	13	41
	Expected	14.138	13.532	13.33	
	Deviation	-4.138	4.468	-0.33	
	Cell Chi-Square	1.2111	1.4752	0.0082	
	Column percent	14.29	26.87	19.70	
Required	Frequency	1	27	7	35
	Expected	12.069	11.552	11.379	
	Deviation	-11.07	15.448	-4.379	
	Cell Chi-Square	10.152	20.659	1.6854	
	Column percent	1.43	40.30	10.61	
Total		70	67	66	203

5.1.3. Video-Essay

Performance of the Video-Essay section appeared to differ according to group membership. Table 5.7. is the result of chi-square analyses on the Video-Essay section. Grouping by both native language and learning experience resulted in significant Chi-square values. When the subjects were grouped by their native language, the Chi-square value was 20.987, which was larger than 13.707 when the subjects were grouped by their learning experience.

When different language groups were compared, the Indo-European group turned out to be the best group, and the Altaic group was slightly better than the Sino-Tibetan group as in Table 5.8. A larger proportion of the IE subjects got

grade 4 than the other two groups. As the cell Chi-square statistic indicated, most of the contribution to the Chi-square value came from those who got grade 4. More subjects among the IE group got grade 4 than the expected frequency, whereas a smaller number of ST subjects got grade 4 than the expected number of students.

Table 5.7. Chi-square analyses on the Video-Essay section

Grouping	DF	Chi-square	Prob
Language Family	4	20.987	.000
Learning Experience	4	13.707	.008

Table 5.8. Cell chi-square analysis when the examinees were grouped by language family

Grade	Statistic	IE	ST	Altaic	Total
2	Frequency	5	4	5	14
	Expected	4.8756	4.5274	4.597	
	Deviation	.1244	-.527	.403	
	Cell Chi-Square	.0032	.0614	.0353	
	Column percent	7.14	6.15	7.58	
3	Frequency	41	57	53	151
	Expected	52.587	48.831	49.582	
	Deviation	-11.59	8.1692	3.4179	
	Cell Chi-Square	2.5531	1.3667	.2356	
	Column percent	58.57	87.69	80.30	
4	Frequency	24	4	8	36
	Expected	12.537	11.642	11.021	
	Deviation	11.463	-7.642	-3.821	
	Cell Chi-Square	10.48	5.0162	1.235	
	Column percent	34.29	6.15	12.12	
Total		70	65	66	201[13]

[13] Two subjects were not encoded in the UIUC EPT data base. Both of them belong to the grammar-and-reading-focused group and also to the Sino-Tibetan group.

Table 5.9. Cell chi-square analysis when the examinees were grouped by learning experience

Grade	Statistic	GR	CO	COMM	Total
2	Frequency	4	5	5	14
	Expected	4.9453	3.1343	5.9204	
	Deviation	-.945	1.8657	-0.92	
	Cell Chi-Square	.1807	1.1105	.1431	
	Column percent	5.63	11.11	5.88	
3	Frequency	62	33	56	151
	Expected	53.338	33.806	63.856	
	Deviation	8.6617	-.806	-7.856	
	Cell Chi-Square	1.4066	.0192	.9664	
	Column percent	87.32	73.33		
4	Frequency	5	7	24	36
	Expected	12.716	.0597	15.224	
	Deviation	-7.716	-1.06	8.7761	
	Cell Chi-Square	4.6824	.1393	5.0592	
	Column percent	7.04	15.56	28.24	
Total		71	45	85	201[13]

When different learning experience groups were compared, the communication-focused learning experience group performed best among three groups. The two biggest aberrations from the expected frequency occurred among those who got grade 4. The grammar and reading-focused group had fewer students at grade 4 than the expected number, whereas the communication-focused group had more students of grade 4 than the expected. The cell Chi-squares for the grammar and reading-focused group and the communication-focused group were 4.6824, and 5.0592, respectively.

5.2. Dimensionality and DIF in the Structure Section

The result of the DIMTEST showed the data of the structure section were essentially unidimensional. As in

Table 5.10, The DIMTEST *T* statistic did not reject the null hypothesis that the data were essentially unidimensional.

Table 5.10. DIMTEST statistic

T	P-value[14]
.906478	.182341

In other words, the structure session measured essentially unidimensional ability (one dominant ability which can be termed structural knowledge).

The SIBTEST detects the items which may form minor dimensions. As explained in 4.3, these items can be either biased items or DIF items depending on whether they can be validated. The present study checked the content-related validity of the flagged items from the SIBTEST to examine whether the items were just functioning differentially or biased against one group, and it found no evidence of disproving the validity of the flagged items[15]. In other words, the flagged items from the SIBTEST were DIF items, not biased items.

The SIBTEST was applied to 3 pairs of different language groups and 3 pairs of different learning experience groups. The SIBTEST produced three types of output: (a) individual DIF, group DIF, and DTF. Table 5.11 and 5.12 are the summaries of the six pair comparisons[16]. DIF items in the

[14] For details of the DIMTEST statistics, see Nandakumur (1991; 1993), Stout (1987), and Stout, Douglas, Junker, and Roussos (1993).
[15] Content-related validity checking was based on my own judgment.
[16] Complete output is in Appendix B.

tables appeared to favor one group over the other significantly at the significance level of .05. Items in the parentheses are group DIF items which appeared to function differentially when they were tested together. Both Beta-uni and SIB-uni are estimators of collective amount of differential functioning. The sign of the estimators indicates which group is favored. If the sign is positive, the reference group (which is in the first line of each pair) is favored, and if the sign is negative, the focal group (which is the second line) is favored.

Table 5.11 shows that three pairs were matched according to language family; Indo-European (IE) vs. Altaic (AL), Altaic vs. Sino-Tibetan (ST), and Sino-Tibetan vs. Indo-European. In comparing the IE group with the AL group, 11 DIF items were detected. Six items of those appeared favorable for the Altaic group and 3 of the rest 5 items appeared for the Indo-European group and two items (item 33, 34) were group DIF items for the IE group. DTF was detected to be favorable for the IE group.

In comparing the ST group and the Altaic group, 7 items functioned differentially. Three items appeared favorable for the ST group and four items for the Altaic group. Group DIF items were not found. DTF was not detected. Effects of the ST group favored items and the Altaic favored items were canceled out.

Nine items were found to function differentially between the Indo-European group and the Sino-Tibetan group. Seven of

them were favorable for the Indo-European group and two items favorable for the Sino-Tibetan group. Group DIF items were not found. DTF appeared very significantly favorable for the IE group.

As shown in Table 5.12, three pairs were also made when the examinees were grouped by their learning experience; grammar-and-reading-focused group (GR) vs. controlled-oral-language-focused group (CL), controlled-oral- language-focused group vs. communication-focused group(COMM), and communication-focused group vs. grammar-and-reading-focused group. In comparing the GR group and the CL group, six DIF items were found. Five items appeared favorable for GR group and one item for the CL group. DTF was canceled out.

Six DIF items were found in comparing the GR group and COMM group. Four items appeared favorable for the GR group and three of those were group DIF items; the amount of each item's bias was not great enough to be considered as a DIF items, but bias was amplified when three items functioned together. Two items were favorable for the COMM group. DTF was canceled out.

Four items were found to be DIF items in comparing the CL group and the COMM group. One item appeared favorable for the CL group and three for the COMM group. As in other comparisons, DTF was canceled out.

Table 5.11. Comparisons between language family groups.

Groups	DIF items	DTF		
AL favored items	6, 19, 24, 28, 29, 32	Beta-uni	SIB -uni	p-value
IE favored items	1, 11, 15, (33, 34)	-5.57	-2.262	*.024
ST favored items	11, 22, 8	Beta-uni	SIB -uni	p-value
AL favored items	2, 14, 32, 44	-2.75	-1.485	.138
IE favored items	1, 2, 15, 20, 40, 44, 14	Beta-uni	SIB -uni	p-value
ST favored items	22, 28	1.811	7.190	*.000

Table 5.12. Comparisons between learning experience groups.

Groups	DIF items	DTF		
GR favored items	3, 6, 8, 36, 48	Beta-uni	SIB -uni	p-value
CL favored items	33	1.89	.837	.402
GR favored items	29, (25, 28, 30)	Beta-uni	SIB -uni	p-value
COMM favored items	15, 33	.119	.626	.531
CL favored items	31	Beta-uni	SIB -uni	p-value
COMM favored items	4, 36, 6	-2.51	-1.225	.221

CHAPTER VI. DISCUSSION

Answer to the research question A: Are language proficiency profiles different across different language groups?

Language proficiency profiles appeared to be different across different language groups. Table 6.1 is the summary of the relative proficiency profiles of three language groups. As in other studies (Kunnan, 1990; Ryan & Bachman, 1992; Alderman & Holland, 1980), the Indo-European group performed equal to or superior to other groups. The results, however, did not exactly conform to expectations. Even though the Indo-European languages have similar structures to that of English, the performance of the IE group on the structure section was not different from others. The IE group, however, performed best both on the pronunciation section and the video-essay section. The Altaic group performed better on the pronunciation section than the Sino-Tibetan group. The performance on video-essay of two groups was not different.

Table 6.1. Relative language proficiency profiles of three language groups (note that this is a descriptive non-inferential analysis).

	Rank of Language Groups		
	1st	2nd	3rd
Structure	No difference		
Pronunciation	IE	AL	ST
Video-Essay	IE	AL = ST	

Equal performance of three groups on the structure section is different from expectation and findings of other studies (Ryan & Bachman, 1992; Alderman & Holland, 1981). This may be due to the homogeneity of subjects' English proficiency. Their proficiency ranges above a certain level but not to a very advanced level. This may imply that effects of linguistic affinity of a language to a target language are not noticeable above a certain level when it comes to structural knowledge.

The Altaic group performed better than the Sino-Tibetan group on the pronunciation section, but not on the essay-writing. It was suspected that the AL group might have more test takers who had studied under the controlled-oral-language-focused instructional method. However, the distributions of two groups across different learning experience appeared about the same (Table 4.4).

Unlike the performance on the other sections, performance on the video-essay section may be less affected by linguistic affinity. It may be more affected by what type of instruction they had. A Chi-square statistic supported this (Table 6.2). The IE group had more students from the COMM and fewer students from the GR compared to the other two groups ($p < .01$). The COMM group was the only group to which writing skill had been emphasized.

Table 6.2. Test of association between native language groups and learning experience groups

	Statistics	GR	CL	COMM	Total
IE	Frequency	13	16	41	70
	Expected	25.172	15.517	29.31	
	Deviation	-12.17	.4828	11.69	
	Cell Chi-Square	5.8861	.015	4.6621	
	Row Percent	18.57	22.86	58.57	
	Column percent	17.81	36.56	48.24	
ST	Frequency	28	15	24	67
	Expected	24.094	14.852	28.054	
	Deviation	3.9064	.1478	-4.054	
	Cell Chi-Square	.6334	.0015	.5859	
	Row Percent	41.79	22.39	35.82	
	Column percent	38.36	33.33	28.24	
AL	Frequency	32	14	20	66
	Expected	23.734	14.631	27.635	
	Deviation	8.266	-.631	-7.635	
	Cell Chi-Square	2.8789	.0272	2.1096	
	Row Percent	48.48	6.90	30.30	
	Column percent	43.84	31.11	23.53	
Total		73	45	85	203

Statistic	DF	Value	Prob
Chi-Square	4	16.800	0.002

Answer to research question B: Are language proficiency profiles different across different learning experience groups?

Differences in learning experience were not necessarily reflected in performance. The three learning experience groups showed differences only on the video-essay section, as seen in Table 6.3 The communication-focused group performed better than the grammar-and-reading-focused group and controlled-oral-language-focused group. This may reflect that the COMM group had more writing instruction than the other

groups. As explained in the previous section, however, learning experience effects on the video-essay section were confounded with first language effects. The COMM group had more IE students than any other groups.

The result of the grammar section comparison can be seen from the same perspective as Landolfi's finding (1991) that instructional method did not cause differences in the achievement of structural knowledge in the long run even though Landolfi's studies were in a different setting. The GR group did not surpass the other groups even though they had grammar and reading focused instruction. Likewise, learning experience did not cause differences in performance on the pronunciation section.

Table 6.3. Relative language proficiency profiles of three learning experience groups (note that this is a descriptive non-inferential analysis

	Rank of Learning Experience Groups		
	1st	2nd	3rd
Structure	No difference		
Pronunciation	No difference		
Video-Essay	COMM	GR = CO	

In the following discussion on research question C and D, only DIF was considered because the contents of all the flagged items were valid measures of structural knowledge.

Answer to research question C: What types of structure test items function differentially across different language groups or are biased against one language group?

Possible reasons to account for the DIF items were sought. However, explaining all DIF items was not possible. Some DIF items had common points while others didn't. Sample DIF items are given in Table 6.4.

The IE group seemed to have a trouble with word orders of *subject + verb* within a subordinate clause (two items including S1[17] in Table 6.4). They also showed weakness in idiomatic expressions such as *can not help + verb-ing* (item S2), and *be used to + ing*. However, they were better than the other groups in choosing appropriate wh-words (S3). A possible reason for poor performance on word order problems might be the flexible word order in many Indo-European language such as Italian and Spanish.

The ST group seemed to have difficulty with sentence connectors in two items and long verb phrases with modals such as *could have seen* (S4), *must have been*, and *will leave*. The AL group did not show clear patterns of strengths and weaknesses.

The collective amount of the DIF, that is, DTF appeared favorable for the IE group against the AL group and the ST group. DTF was canceled out between the AL group and the ST group.

[17] Only a few items are provided as examples. The items are renumbered due to a security reason.

Table 6.4. Sample DIF items when the language groups are compared.

S1. "Linda knows when the boys are leaving."
"Did she say where _____ ?"

a. were they going
b. they going were
c. were going they
d. they were going

S2. Martin can't help _____ sorry for himself.

a. to be felt
b. to feel
c. feeling
d. that he feels

S3. "I don't know what Mary is going to do with all those clothes."
"And I wonder _____ she is going to wear them."

a. which
b. where
c. what
d. that

S4. If I had not missed the bus, I _____ them before they left.

a. should see
b. could see
c. should have seen
d. could have seen

Answer to research question D: What types of structure test items function differentially across different learning experience groups or are biased against one learning experience group?

It does not seem to be possible to provide plausible accounts for DIF items from the comparisons of learning experience groups. The CL group did worst in item S5 where

the format is conversation style which the CL group was expected to be good at. The GR group did worse in item S6 which seemed to be very easy if they had intensive grammatical instruction (for + noun and to + bare infinitive to express the meaning of objective). Compared to the COMM group, the GR group performed better in items with tense problems in the subordinate clause (two items including S7 in Table 6.5). This may be because meaning was more focused to

Table 6.5. Sample DIF items when the learning experience groups are compared.

S5. "How long have you lived in this town?"
"I _____ here for six years by next week."

a. would live
b. would have lived
c. will have lived
d. will live

S6. Ms. Peters went to the hardware store _____ some paint.

a. for buy
b. for
c. for to buy
d. for buying

S7. Students will do well on the history test if they _____ most of answers.

a. will know
b. had known
c. are knowing
d. know

S8. Jane rode her bike _____ the street.

a. from
b. over
c. up
d. at

---

the COMM group. If they chose the answer based on the meaning, it would be natural for them to choose the future tense where a present tense was required. In item S1 where the GR group was favored over the COMM group (see Table 5.12), meaning did not matter in choosing the answer, as well. Inversely, the COMM group performed better than the other groups where meaning mattered in three items including S8, but they had difficulties where meaning was not crucial for choosing the answer in five items such as S1 and S7.

In all comparisons of three pairs, DTF was not significant because the collective amount of DIF items was canceled out.

Answer to research question E: Which of the two characteristics can explain test takers' performance on the UIUC EPT better?

Native language appeared more influential on the performance on the UIUC EPT than learning experience. That is, there were more differences in the performance on the UIUC EPT when test takers were grouped according to native language than when they were grouped according to learning experience. Grouping by both native language and learning experience did not result in group differences in the structure section. Three language groups were different in both the pronunciation section and the video-essay section whereas three learning experience group differed only in the video-essay section.

Even though both ways of grouping did not result in group differences on the structure section, the item level comparison showed that test takers performed more differentially when they were grouped according to native language. The collective amount of DIF items, DTF, appeared significantly differential only when three language groups were compared. The DTF was not significant among three learning experience groups because the collective amount of DIF was canceled out.

These findings suggest that test takers' native languages be given more weight than learning experience in understanding the problematic areas of students and designing ESL classes if variables other than English proficiency are included in the EPT administration. One way of implementing this is using a program 'FACET'. Based on multi-faceted Rasch model, this program can examine whether any facet, in this case, native language and learning experience, has differential effects across groups. These differential effects can be statistically adjusted (A. Liu, personal communication, March, 1997)

CHAPTER VII. CONCLUSION

The main purpose of EPTs is to diagnose the problematic areas and to place students into appropriate ESL classes. The question is whether test scores are enough to make valid inferences and to determine the right treatment. There is no doubt in that the test score is the most important indicator of language proficiency. However, test performance is the composite of various factors as well as language ability. That is, those factors may also have to be considered in inference making processes so that students may get the better help. In this vein studying the effects of test taker characteristics help to make valid use of EPTs.

This study examined and compared the effects of two test taker characteristics, native language and learning experience, on performance on the UIUC EPT. This study was intended to help understand test performance on the UIUC EPT and provide a basis for fair interpretation and use of test scores.

Learning experience effects were less present compared to native language effects. At the comparisons of overall proficiency profiles, learning effects were present only at the video-essay section. The COMM group performed better than the other two groups. This result was confounded with language effects. Compared to other language groups, the IE group had more subjects of the COMM group and less subjects of the GR group. At the item level analysis of the structure section, DIF appeared stronger when test takers were compared

according to their native language than their learning experience.

This study could provide a positive answer to some researchers' suspicion that different learning experiences may be responsible for differences in performance (Fahardy, 1982; Kunnan, 1990). Learning experience effects were present in some of structure items and the video-essay section, though they were not as global and strong as the language effects.

This results of this study need to be replicated with more subjects. The DIF analyses of this study may have large Type I errors due to the small numbers of subjects per group. Small sample sizes also resulted in information losses when minor groups were removed from the analyses and when the final grades (Pass, Recommended, and Required) for the pronunciation section were used for analyses instead of original metrics (see p.28) used by graders.

The findings of this study should also be supplemented by a close examination of the subjects' performance in ESL classes. This study could not analyze the details of the performance on the pronunciation section and the video-essay section because only quantitative data were available. A close examination of the subjects' performance in ESL classes could not only confirm the findings of this study but also pinpoint problematic areas a particular group would have. This information can also be reflected in the revision of the test.

Finally, the results of this study is limited to only certain groups of examinees. The EPT takers had enough English proficiency to get admitted to the University of Illinois at Urbana-Champaign, but their proficiency was not enough to get exempt from the EPT. If the same study were done with learners having a wider range of proficiency, the results would be different since beginner level learners seem to rely more on sources other than the target language in producing the target language (see Krashen & Terrell, 1983).

REFERENCES

Alderman, D. L., & Holland, P. W. (1981). Item performance across native language groups on the test of English as a foreign language (TOEFL Research Rep. No. 9). Princeton, NJ: Educational Testing Service.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 51, 201-238.

American Psychological Association. (1966). Standards for educational and psychological tests and manuals. Washington, DC: Author.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1974). Standards for educational and psychological tests. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Brown, H. D. (1987). Principles of language learning and teaching (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall Inc.

Brown, J. D. (1983). An exploration of morpheme-group interactions. In K. M. Bailey, M. Long,, & S. Peck, (Eds.), Second language acquisition studies (pp. 25-40). Rowley, MA: Newbury House.

Carroll, J. B. (1961). Fundamental considerations in testing English proficiency of foreign students. In Testing the English Proficiency of Foreign Students (pp. 30-40). Washington, D.C.: Center for Applied Linguistics.

Celce-Murcia, M. (1991). (Ed.). Teaching English as a second or foreign language (2nd ed.). Boston, MA: Heinle & Heinle Publishers.

Diller, K. C. (1971). Generative Grammar, structural Linguistics, and language teaching. Rowley, MA: Newbury House Publishers

Dulay, H., Burt, M. & Krashen, S. (1982). Language Two. New York: Oxford University Press.

Dunbar, S. B. (1982). Construct validity and internal structure of a foreign language test for several native language groups. Paper presented at the annual meeting of the American Educational Research Association, New York.

Eckman, F. (1977). Markedness and the contrastive analysis hypothesis. Language Learning, 27, 315-330.

Farhady, H. (1982). Measures of language proficiency from the learner's perspective. TESOL Quarterly, 16, 43-59.

Grimes, B. F. (1992). (Ed.). Ethnologue: Language of the world (12th ed.). Dallas, Texas: Summer Institutes of Linguistics, Inc.

Harris, D. P., & Palmer, L. A. (1970). A comprehensive English language test for speakers of English as a second language. New York, NY: McGraw-Hill Book Co.

Hosley, D. (1978). Performance difference of foreign students on the TOEFL. TESOL Quarterly, 12, 99-100.

Kellerman, E. (1977). Towards a characterization of the strategies of transfer in second language learning. Interlanguage Studies Bulletin, 2, 58-145.

Kirk, R. E. (1995). Experimental design: Procedures for the behavioral sciences (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Krashen, S. D., & Terrell, T. D. (1983). The natural approach. Haywood, CA: The Alemany Press.

Krashen, S. (1985). The input hypothesis: Issues and implication. Longman, New York.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. TESOL Quarterly, 24, 741-746.

Landolfi, L. (1991). A macro-study on the impact of "method" on second language acquisition. Rassegna Italiana di Linguistica Applicata, 22, 99-112.

Larsen-Freeman, D., & Long, M. (1991). Foreign language education: Towards a national research agenda. Washington, D.C.: National Foreign Language Center

Long, M. (1988). Instructed interlanguage development. In L. Beebe (Ed.), Issues in Second Language Acquisition: Multiple perspectives (pp. 115-141). New York: Newbury House.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement. (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Mitchell, L. A. (1991). English placement testing: Recognizing the sociolinguistic context. World Englishes, 10, 75-82.

Moriyama, Y. (1994). Construct validation of the United Nations association test of English level a test and its implications for further development of the test. Unpublished master's thesis, University of Illinois at Urbana-Champaign.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. Journal of Educational Measurement, 28, 99-117.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. Applied Psychological Measurement, 17, 29-38.

Oltman, P. K., Stricker, L. J., & Barrows, T. (1988). Native Language, English Proficiency, and the Structure of the Test of English as a Foreign Language (TOEFL Research Rep. No. 6). Princeton, NJ: ETS.

Perkins, K., & Larsen-Freeman, D. (1975). The effect of formal language instruction on the order of morpheme acquisition. Language Learning, 25, 237-243.

Plaister, T. H., & Blatchford, C. H.(1971). Plaister aural comprehension test (PACT). Honolulu: University of Hawaii, English Language Institute.

Politzer, R. L., & McGroarty. (1985). An exploratory study of learning behavior and their relationship to gains in linguistic and communicative competence. TESOL Quarterly, 19, 103-123.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. Journal of Educational Measurement, 33, 215-230.

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. Language Testing, 9, 12-29.

Sawatdirakpong, S. (1993). Native languages and differential dimensionality of English as a second language test proficiency: An exploratory study. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), Review of Research in Education 19 (pp. 405-450). Washington, DC: AERA.

Spuring, S., & Ilyin, D. (1985). The impact of learner variables on language test performance. TESOL Quarterly, 19, 283-299.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrica, 52, 589-617.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). Dimetest manual. Unpublished manuscript, University of Illinois at Urbana-Champaign.

Stout, W., & Roussos, L. (1992). Sibtest user manual. Unpublished manuscript, University of Illinois at Urbana-Champaign.

Stout, W., & Shealy, R. (1992). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Unpublished manuscript, University of Illinois at Urbana-Champaign.

Swinton, S. S., & Powers, D. E. (1980). Factor analysis of the test of English as a foreign language for several language groups (TOEFL Research Rep. No. 6). Princeton, NJ: ETS.

U.S. Bureau of the Census (1994). Statistical abstract of the United States: 1994 (114th ed.). Washington, DC: Author.

Wode, H. (1981). Language-acquisitional universals: a unified view of language acquisition. In H. Winitz (Ed.) Native language and foreign language acquisition (pp. 218-234). Annals of the New York Academy of Sciences.

Zobl, H. (1982). A direction for contrastive analysis: The comparative study of developmental sequences. TESOL Quarterly, 16, 169-83.

APPENDIX A

EPT INFORMATION FORM

[This form was designed and administered for the present study and is not part of the operational UIUC EPT]

This questionnaire is designed to study what types of classroom experience you have had in your country. Thank you for taking the time to answer.

Family Name _____ Given Name _____

Social Security Number: _____-____-_____

Home Country _____ First Language _____

Other language you speak, besides English _____

UIUC Department _____

Status: undergraduate graduate visiting scholar

These questions are about the institutes where you studied most of your English.

1. Did you have textbooks for your English classes? Yes No

   1.1. If yes, how much did your English teachers rely on the textbooks? (Did you follow through the textbooks?) Check one

   ( ) never did anything other than the textbooks.

   ( ) combined 75% of the textbooks and other activities

   ( ) combined 50% of the textbooks and other activities

   ( ) combined 25% of the textbooks and other activities

1.2. Can you recall the percentage of each part of your **textbooks**?

Reading (%) Grammar (%) Writing (%)

Speaking (%) Listening (%) Other _____ (%)

1.3. Can you recall the percentage of each **activity** in your class?

Reading (%) Grammar (%) Writing (%)

Speaking (%) Listening (%) Other _____ (%)

2. Were you allowed to use your native language in your English classes?

Yes **No**

If **yes**, how often? (circle one)

rarely often always

3. Did your teacher speak or explain things in English in your English classes

Yes **No**

If **yes**, how often? (circle one)

rarely often always

4. Did you or your teachers translate your textbooks sentence by sentence?

Yes **No**

5. Did you have language lab facilities? If so, how often did you use the lab?

( ) 1. very rarely (once a semester)
( ) 2. rarely (two or three times a semester
( ) 3. often (once or twice a month)
( ) 4. very often (more than once a week)

6. Did your teacher ask you to memorize the dialogue in your textbooks?

Yes **No**

7. Can you rank the following language skills from the one that was regarded most important? **1st (most) - 4th (least)**

Reading ( ) Writing ( ) Speaking ( ) Listening ( )

8. How were the percentages of the following parts reflected in your final grades for your English classes?

Reading tests (%) Writing tests (%) Speaking tests (%)

Grammar tests (%) Listening tests (%)

9. Were you often asked to speak in English in your class? **Yes** **No**
(Speaking does not mean repetition of the teachers' words)

How often? (circle one) **never** **rarely** **often** **always**

10. Were you always required to use correct English? **Yes** **No**
How did your teachers try to correct you?

( ) Whenever you made mistakes.
( ) Sometimes they did, but other times they didn't.
( ) They never did.

11. Could most of your English teachers speak English fluently? **Yes** **No**

APPENDIX B

SIB TEST OUTPUT

1. Grammar-and-Reading vs. Controlled-Oral-Language

Reference group = Grammar-and-reading-focused group
Focal group = Controlled-oral-language-focused group

First run: All items were tested individually.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
1	-.056	-.515	.606 E	1.20	.274 E	1.47
2	-.008	-.080	.936 E	.13	.718 E	.83
3	.140	3.674	.000 E	3.11	.078 E	*****
4	.053	.663	.507 E	.90	.342 E	-2.19
5	.021	.282	.778 E	.11	.735 E	-.89
6	.220	2.778	.005 E	3.18	.074 E	-3.28
7	.005	.050	.960 E	.09	.765 E	.55
8	.221	2.089	.037 E	2.85	.091 E	-2.11
9	.149	1.294	.196 E	.57	.451 E	-.95
10	.127	2.429	.015 E	.29	.590 E	*****
11	.017	.161	.872 E	.91	.341 E	1.29
12	-.035	-1.160	.246 E	.63	.427 E	3.60
13	-.028	-.537	.591 E	1.07	.301 E	2.78
14	.069	.768	.442 E	.05	.820 E	-.04
15	-.059	-.526	.599 E	3.71	.054 E	2.73
16	-.043	-.456	.648 E	.00	.954 E	.33
17	.083	.792	.428 E	1.10	.294 E	-1.38
18	.102	3.226	.001 E	3.17	.075 E	*****
19	-.052	-.795	.427 E	.01	.941 E	.64
20	-.131	-1.400	.161 E	1.84	.175 E	1.54
21	.009	.386	.699 E	.00	.989 E	-1.42
22	.152	1.217	.224 E	2.12	.145 E	-1.96
23	.050	2.788	.005 E	.00	1.000 E	-1.63
24	.047	.391	.696 E	.00	.954 E	-.28
25	.069	1.308	.191 E	.56	.453 E	.00
26	.124	1.933	.053 E	.05	.820 E	-.38
27	.053	.914	.360 E	.88	.347 E	-2.94
28	.171	2.908	.004 E	.62	.429 E	-2.45
29	.161	1.424	.154 E	.02	.899 E	-.38
30	.098	1.806	.071 E	.04	.842 E	1.14
31	-.040	-.436	.662 E	1.09	.297 E	1.90
32	.093	1.187	.235 E	.03	.852 E	-.64
33	-.376	-3.590	.000 E	12.52	.000 E	4.10
34	-.193	-1.798	.072 E	6.72	.010 E	4.08
35	.160	1.212	.225 E	.95	.330 E	-1.24
36	.335	3.280	.001 E	6.08	.014 E	-3.22
37	-.113	-1.207	.227 E	1.21	.271 E	1.28
38	.073	2.398	.017 E	3.16	.075 E	*****
39	.143	1.221	.222 E	.00	.960 E	-.21

40	.037	.525	.600 E	.01	.940 E	-.34
41	.005	.068	.946 E	.45	.501 E	1.55
42	.100	1.033	.302 E	.62	.429 E	-1.04
43	.139	2.003	.045 E	.02	.899 E	.92
44	-.034	-.298	.766 E	.80	.371 E	1.28
45	.010	.094	.925 E	.00	.984 E	-.27
46	.151	2.541	.011 E	1.97	.160 E	-2.69
47	.122	1.990	.047 E	.00	.975 E	-.36
48	-.016	-.116	.908 E	.55	.456 E	.95
49	.090	1.582	.114 E	.04	.847 E	-.48
50	.034	-99.000	.000 E	.21	.646 E	-.21

Reference group = Grammar-and-reading-focused group
Focal group = Controlled-oral-language-focused group

Second run

The unflagged items from the first run entered into the test

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
1	-.134	-1.117	.264 E	1.06	.303 E	1.21
2	.004	.044	.965 E	.00	.984 E	.36
4	.009	.147	.883 E	.14	.711 E	-1.13
5	.039	.512	.608 E	.01	.908 E	-.24
7	.004	.046	.963 E	.14	.711 E	.59
9	-.013	-.121	.904 E	.00	.965 E	-.25
11	-.046	-.383	.702 E	.02	.876 E	.36
12	-.062	-1.346	.178 E	.01	.923 E	1.84
13	-.055	-1.150	.250 E	.36	.548 E	1.98
14	-.054	-.554	.579 E	.11	.744 E	.66
15	-.170	-1.724	.085 E	1.01	.314 E	1.38
16	-.049	-.826	.409 E	.05	.815 E	-.04
17	.116	1.082	.279 E	.40	.528 E	-.70
19	-.089	-1.349	.177 E	.11	.741 E	-.02
20	-.151	-1.311	.190 E	2.29	.131 E	1.73
21	-.051	-1.537	.124 E	.01	.937 E	1.09
22	.156	1.322	.186 E	.77	.380 E	-1.08
24	.063	.622	.534 E	.00	.961 E	.18
25	-.033	-.952	.341 E	.41	.522 E	.00
26	.029	.454	.650 E	.01	.918 E	-1.22
27	.105	1.575	.115 E	2.14	.143 E	-4.59
29	.102	.805	.421 E	.00	.976 E	-.15
30	.050	1.143	.253 E	.03	.868 E	-.51
31	-.045	-.559	.576 E	.10	.753 E	.60
32	-.025	-.303	.762 E	.08	.780 E	.00
35	.239	2.336	.020 E	2.53	.112 E	-1.82
37	-.085	-.879	.380 E	.46	.496 E	.93
39	.115	1.172	.241 E	1.17	.279 E	-1.45
40	-.060	-.689	.491 E	.05	.820 E	.09
41	-.059	-.894	.371 E	.00	.996 E	.57
42	.166	1.639	.101 E	1.57	.210 E	-1.58
44	-.067	-.637	.524 E	.26	.613 E	.83
45	.029	.228	.820 E	.31	.578 E	-.72
48	-.206	-2.123	.034 E	.05	.827 E	.40
49	-.011	-.289	.772 E	.00	.961 E	-.77

Reference group = Grammar-and-reading-focused group
Focal group = Controlled-oral-language-focused group

Third run:

All flagged items were tested against the valid items which were not flagged from the second run.

Valid subtest items:

1	2	4	5	7	9	11	12	13	14
15	16	17	19	20	21	22	24	25	26
27	29	30	31	32	35	37	39	40	41
42	44	45	49						

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
3	.111	2.491	.013 E	2.22	.136 E	-5.17
6	.209	1.985	.047 E	1.95	.162 E	-2.28
8	.247	2.057	.040 E	2.63	.105 E	-1.98
18	.049	1.491	.136 E	.39	.535 E	*****
33	-.402	-4.135	.000 E	11.38	.001 E	4.00
34	-.116	-1.030	.303 E	4.89	.027 E	2.49
36	.184	1.813	.070 E	3.44	.064 E	-1.98
38	.042	.791	.429 E	2.37	.124 E	-5.74
50	.015	.392	.695 E	.01	.934 E	-1.29
10	.062	1.634	.102 E	.24	.626 E	-3.08
23	-.018	-.495	.621 E	.09	.768 E	-.48
28	.071	1.170	.242 E	1.57	.210 E	-2.27
46	.070	1.235	.217 E	.88	.349 E	-1.70
47	.113	1.310	.190 E	.06	.804 E	.02
43	.087	1.343	.179 E	.01	.916 E	-.80
48	-.206	-2.123	.034 E	.05	.827 E	.40

```
Reference group = Grammar-and-reading-focused group
Focal group = Controlled-oral-language-focused group

Fourth run: Tests of group DIF and DTF

Reference group favored items: 3, 6, 8, 36, 48
Focal group favored items: 33


-----------------------------------------------------------------
 OUTPUT FOR RUN NUMBER   1          OUTPUT FOR RUN NUMBER   1

 Suspect subtest items:
   3   6   8  48  36

 Valid subtest items:
   1   2   4   5   7   9  10  11  12  13
  14  15  16  17  18  19  20  21  22  23
  24  25  26  27  28  29  30  31  32  33
  34  35  37  38  39  40  41  42  43  44
  45  46  47  49  50
 proportion of Ref. grp. examinees eliminated =    .224
 proportion of Focal grp. examinees eliminated =   .255

                          SIB-uni                Mantel-Haenszel Results
                        p-value for                    p-value for
              SIB-uni   DTF against                    DIF against
                 z      either Ref.        Chi        either Ref.   Delta
 Beta-uni statistic  or Foc. grp.          sqr.    or Foc. grp. (D-DIF)
    .541     3.003        .003

-----------------------------------------------------------------
 CUTPUT FOR RUN NUMBER   2          OUTPUT FOR RUN NUMBER   2

 Suspect subtest items:
  33

 Valid subtest items:
   1   2   3   4   5   6   7   8   9  10
  11  12  13  14  15  16  17  18  19  20
  21  22  23  24  25  26  27  28  29  30
  31  32  34  35  36  37  38  39  40  41
  42  43  44  45  46  47  48  49  50
 proportion of Ref. grp. examinees eliminated =    .211
 proportion of Focal grp. examinees eliminated =   .277

                          SIB-uni                Mantel-Haenszel Results
                        p-value for                    p-value for
              SIB-uni   DIF against                    DIF against
                 z      either Ref.        Chi        either Ref.   Delta
 Beta-uni statistic  or Foc. grp.          sqr.    or Foc. grp. (D-DIF)
   -.376    -3.590        .000           12.52         .000        4.096
```

OUTPUT FOR RUN NUMBER 3 OUTPUT FOR RUN NUMBER 3

Suspect subtest items:

3	6	8	48	36	33

Valid subtest items:

1	2	4	5	7	9	10	11	12	13
14	15	16	17	18	19	20	21	22	23
24	25	26	27	28	29	30	31	32	34
35	37	38	39	40	41	42	43	44	45
46	47	49	50						

proportion of Ref. grp. examinees eliminated = .197
proportion of Focal grp. examinees eliminated = .234

		SIB-uni p-value for DTF against either Ref. or Foc. grp.	Mantel-Haenszel Results		
	SIB-uni z			p-value for DIF against either Ref. or Foc. grp.	
Beta-uni	statistic		Chi sqr.		Delta (D-DIF)
.189	.837	.402			

2. Grammar-and-Reading vs. Communication

Reference group = Grammar-and-reading-focused group
Focal group = Communication-focused group

First run:

All items were tested individually.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value		Mantel-Haenszel Chi sqr.	p value		Delta (D-DIF)
1	-.123	-1.459	.145	E	3.09	.079	E	1.63
2	-.072	-.903	.367	E	.07	.791	E	.48
3	.040	1.331	.183	E	.93	.335	E	-4.04
4	-.099	-1.336	.181	E	2.39	.122	E	1.86
5	-.072	-1.366	.172	E	.54	.463	E	1.69
6	.007	.145	.885	E	.00	.963	E	-.39
7	-.005	-.055	.956	E	1.06	.303	E	-1.11
8	.000	.002	.999	E	.06	.800	E	-.42
9	.079	.912	.362	E	.07	.789	E	-.36
10	.025	1.109	.268	E	1.43	.232	E	*****
11	-.120	-1.458	.145	E	2.02	.155	E	1.40
12	-.001	-.031	.976	E	.00	.967	E	.77
13	-.030	-.529	.597	E	.06	.807	E	.81
14	-.079	-1.033	.302	E	.07	.792	E	.47
15	-.263	-3.008	.003	E	7.85	.005	E	2.89
16	.025	.376	.707	E	.16	.687	E	-.71
17	-.004	-.045	.964	E	.11	.738	E	-.49
18	.002	.136	.892	E	1.04	.307	E	*****
19	-.030	-.582	.560	E	.47	.493	E	1.34
20	-.154	-1.894	.058	E	.89	.346	E	1.02
21	.006	.169	.866	E	.08	.775	E	1.60
22	.130	1.635	.102	E	2.48	.116	E	-1.60
23	.022	.523	.601	E	.00	.947	E	-.94
24	.032	.429	.668	E	.05	.824	E	-.39
25	.086	1.929	.054	E	2.17	.141	E	-3.45
26	-.011	-.249	.803	E	.14	.706	E	1.50
27	.036	.799	.424	E	.04	.836	E	-1.34
28	.151	2.634	.008	E	2.08	.150	E	-2.84
29	.191	2.373	.018	E	3.41	.065	E	-1.68
30	.071	1.447	.148	E	2.01	.157	E	-2.52
31	.022	.268	.789	E	.01	.926	E	-.09
32	.034	.536	.592	E	1.95	.163	E	-1.67
33	-.265	-3.028	.002	E	5.81	.016	E	2.36
34	-.146	-1.775	.076	E	.63	.426	E	.86
35	-.033	-.364	.716	E	.01	.920	E	.28
36	-.007	-.084	.933	E	.03	.861	E	.03
37	-.089	-1.056	.291	E	.03	.874	E	.02
38	.013	.417	.677	E	.61	.435	E	-3.17
39	-.040	-.510	.610	E	.01	.931	E	-.15
40	-.087	-1.532	.125	E	.01	.922	E	-.25
41	-.053	-1.066	.286	E	.48	.487	E	1.55
42	-.003	-.043	.965	E	.00	.978	E	.14

43	-.049	-1.130	.258	E	.21	.647	E 1.28
44	-.148	-1.696	.090	E	1.60	.207	E 1.22
45	-.011	-.129	.897	E	.00	.967	E -.15
46	.071	1.240	.215	E	4.65	.031	E -3.93
47	.136	2.010	.044	E	4.43	.035	E -2.77
48	-.008	-.095	.924	E	.00	.960	E .11
49	.032	.774	.439	E	.04	.835	E -1.05
50	.031	1.094	.274	E	.85	.358	E -4.27

Reference group = Grammar-and-reading-focused group
Focal group = Communication-focused group

Second run:
The unflagged items from the first run entered into the test.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value		Mantel-Haenszel Chi sqr.	p value		Delta (D-DIF)
1	-.097	-1.150	.250	E	2.98	.084	E	1.54
2	-.056	-.861	.389	E	.59	.443	E	.97
3	.050	1.764	.078	E	.48	.489	E	-2.84
4	-.108	-1.518	.129	E	3.52	.061	E	2.04
5	-.081	-1.868	.062	E	1.73	.188	E	2.30
6	-.002	-.045	.964	E	.02	.877	E	-.63
7	.057	.755	.450	E	.24	.623	E	-.64
8	.035	.531	.596	E	1.62	.203	E	-1.44
9	-.006	-.082	.934	E	.04	.849	E	-.35
10	.032	1.698	.089	E	1.54	.214	E	*****
11	-.129	-1.581	.114	E	.82	.366	E	.86
12	-.013	-.294	.769	E	.00	.970	E	.48
13	-.014	-.303	.762	E	.27	.601	E	1.31
14	-.023	-.375	.707	E	.03	.862	E	.38
16	-.059	-.904	.366	E	1.06	.304	E	-1.27
17	.053	.624	.533	E	.09	.769	E	-.40
18	.002	.126	.900	E	.56	.453	E	*****
19	-.053	-1.192	.233	E	.91	.340	E	1.53
20	-.126	-1.693	.090	E	.96	.326	E	.99
21	-.008	-.278	.781	E	.06	.805	E	-.24
22	.133	1.604	.109	E	4.06	.044	E	-1.90
23	-.004	-.121	.904	E	.17	.679	E	-1.25
24	-.004	-.056	.955	E	1.26	.261	E	-1.18
25	.109	2.605	.009	E	2.85	.092	E	-3.87
26	-.031	-1.066	.286	E	.36	.548	E	2.19
27	-.007	-.203	.839	E	.01	.932	E	.53
30	.086	1.732	.083	E	1.46	.227	E	-1.99
31	.076	.980	.327	E	.00	.964	E	-.11
32	-.003	-.045	.964	E	1.33	.249	E	-1.29
34	-.069	-.874	.382	E	1.04	.308	E	1.08
35	-.055	-.777	.437	E	.12	.727	E	.47
36	.009	.140	.889	E	.16	.689	E	-.62
37	.006	.074	.941	E	.00	.992	E	.14
38	.026	.928	.354	E	.14	.707	E	-2.61
39	-.047	-.649	.517	E	.00	.971	E	.18
40	-.043	-.860	.390	E	1.24	.266	E	1.86
41	-.058	-1.422	.155	E	2.14	.144	E	2.30
42	-.018	-.219	.827	E	.00	.998	E	-.16
43	-.025	-.540	.589	E	.03	.863	E	.81
44	-.115	-1.647	.100	E	3.37	.066	E	2.05
45	.045	.579	.562	E	.88	.349	E	-.94
46	.090	1.822	.068	E	2.38	.123	E	-2.53
48	.007	.081	.935	E	.00	.948	E	.09
49	.035	.806	.420	E	.28	.595	E	-1.26
50	.020	.661	.508	E	.31	.575	E	-1.99

Reference group = Grammar-and-reading-focused group
Focal group = Communication-focused group

Third run:

All flagged items were tested against the valid items which were not flagged from the second run.

Valid subtest items:

1	2	4	6	7	8	9	11	12	13
14	16	17	18	19	21	22	23	24	26
27	31	32	34	35	36	37	38	39	40
41	42	43	45	48	49	50			

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
15	-.288	-3.625	.000 E	9.82	.002 E	3.08
33	-.212	-2.764	.006 E	5.57	.018 E	2.30
47	.100	1.454	.146 E	2.87	.090 E	-1.97
28	.075	1.654	.098 E	2.02	.155 E	-2.78
29	.179	2.130	.033 E	2.35	.125 E	-1.39
3	.025	.818	.413 E	.64	.423 E	-3.24
5	-.070	-1.452	.147 E	.58	.447 E	1.56
10	.013	.522	.602 E	.72	.396 E	*****
20	-.099	-1.254	.210 E	.23	.634 E	.61
25	.081	1.853	.064 E	3.45	.063 E	-3.84
30	.088	1.779	.075 E	1.78	.182 E	-2.51
44	-.136	-1.928	.054 E	1.28	.258 E	1.24
46	.064	1.189	.235 E	3.57	.059 E	-2.83

Reference group = Grammar-and-reading-focused group
Focal group = Communication-focused group

Fourth run: Tests of suspected items and groups of items from the third run

Referenced group favored group DIF items: 28, 25, 30
Focal group favored items: 44

OUTPUT FOR RUN NUMBER 1 OUTPUT FOR RUN NUMBER 1

Suspect subtest items:
28 25 30

Valid subtest items:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	26	27	29	31	32	33
34	35	36	37	38	39	40	41	42	43
44	45	46	47	48	49	50			

proportion of Ref. grp. examinees eliminated = .118
proportion of Focal grp. examinees eliminated = .163

		SIB-uni p-value for DTF against either Ref.	Mantel-Haenszel Results		
	SIB-uni z			p-value for DIF against	
Beta-uni	statistic	or Foc. grp.	Chi sqr.	either Ref. or Foc. grp.	Delta (D-DIF)
.363	4.595	.000			

OUTPUT FOR RUN NUMBER 2 OUTPUT FOR RUN NUMBER 2

Suspect subtest items:
44

Valid subtest items:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	45	46	47	48	49	50	

proportion of Ref. grp. examinees eliminated = .184
proportion of Focal grp. examinees eliminated = .291

		SIB-uni p-value for DIF against either Ref.	Mantel-Haenszel Results		
	SIB-uni z			p-value for DIF against	
Beta-uni	statistic	or Foc. grp.	Chi sqr.	either Ref. or Foc. grp.	Delta (D-DIF)
-.148	-1.696	.090	1.60	.207	1.221

```
Reference group = Grammar-and-reading-focused group
Focal group = Communication-focused group

Fifth run: Tests of group DIF and DTF

Referenced group favored items: 25, 28, 29, 30
Focal group favored items: 15, 33

-----------------------------------------------------------------
 OUTPUT FOR RUN NUMBER   1          OUTPUT FOR RUN NUMBER   1

 Suspect subtest items:
  29  28  25  30

 Valid subtest items:
   1   2   3   4   5   6   7   8   9  10
  11  12  13  14  15  16  17  18  19  20
  21  22  23  24  26  27  31  32  33  34
  35  36  37  38  39  40  41  42  43  44
  45  46  47  48  49  50
 proportion of Ref. grp. examinees eliminated =    .118
 proportion of Focal grp. examinees eliminated =    .151

                      SIB-uni                Mantel-Haenszel Results
                    p-value for                   p-value for
           SIB-uni  DTF against                   DIF against
             z      either Ref.          Chi     either Ref.   Delta
Beta-uni statistic  or Foc. grp.         sqr.    or Foc. grp. (D-DIF)
   .557    4.753       .000

-----------------------------------------------------------------
 OUTPUT FOR RUN NUMBER   2          OUTPUT FOR RUN NUMBER   2

 Suspect subtest items:
  15  33

 Valid subtest items:
   1   2   3   4   5   6   7   8   9  10
  11  12  13  14  16  17  18  19  20  21
  22  23  24  25  26  27  28  29  30  31
  32  34  35  36  37  38  39  40  41  42
  43  44  45  46  47  48  49  50
 proportion of Ref. grp. examinees eliminated =    .145
 proportion of Focal grp. examinees eliminated =    .244

                      SIB-uni                Mantel-Haenszel Results
                    p-value for                   p-value for
           SIB-uni  DTF against                   DIF against
             z      either Ref.          Chi     either Ref.   Delta
Beta-uni statistic  or Foc. grp.         sqr.    or Foc. grp. (D-DIF)
  -.508   -3.800       .000
```

OUTPUT FOR RUN NUMBER 3 OUTPUT FOR RUN NUMBER 3

Suspect subtest items:
29 28 25 30 15 33

Valid subtest items:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	16	17	18	19	20	21
22	23	24	26	27	31	32	34	35	36
37	38	39	40	41	42	43	44	45	46
47	48	49	50						

proportion of Ref. grp. examinees eliminated = .105
proportion of Focal grp. examinees eliminated = .186

		SIB-uni p-value for DTF against either Ref. or Foc. grp.	Mantel-Haenszel Results p-value for DIF against either Ref. or Foc. grp.	
Beta-uni	SIB-uni z statistic		Chi sqr.	Delta (D-DIF)
.119	.626	.531		

3. Controlled-Oral-Language vs. Communication

Reference group = Controlled-oral-language-focused group
Focal group = Communication-focused group

First run:

All items were tested individually.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value		Mantel-Haenszel Chi sqr.	p value		Delta (D-DIF)
1	-.042	-.368	.713	E	.02	.896	E	.31
2	-.079	-.813	.416	E	.00	.984	E	-.25
3	-.033	-.743	.458	E	1.06	.304	E	2.49
4	-.159	-2.324	.020	E	2.45	.117	E	2.61
5	-.040	-.565	.572	E	2.01	.156	E	3.04
6	-.201	-2.591	.010	E	.48	.488	E	1.20
7	.036	.334	.738	E	2.40	.122	E	-1.86
8	-.123	-1.263	.207	E	3.29	.070	E	2.75
9	.060	.498	.619	E	.02	.897	E	.08
10	-.012	-.238	.812	E	.03	.869	E	1.00
11	-.064	-.628	.530	E	.10	.751	E	-.54
12	.052	1.315	.189	E	.01	.915	E	-.68
13	.055	1.395	.163	E	.04	.838	E	-.29
14	-.097	-1.205	.228	E	.00	.951	E	.20
15	-.149	-1.472	.141	E	1.61	.205	E	1.73
16	.113	1.110	.267	E	.42	.519	E	-1.04
17	.022	.199	.842	E	.00	.975	E	.17
18	-.023	-.552	.581	E	.02	.896	E	1.91
19	-.011	-.234	.815	E	.00	.997	E	.55
20	.085	.918	.359	E	.93	.335	E	-1.10
21	.093	2.100	.036	E	.02	.892	E	1.22
22	-.009	-.081	.935	E	.00	.973	E	-.17
23	.070	1.288	.198	E	.06	.801	E	.43
24	-.055	-.483	.629	E	.03	.859	E	.04
25	.034	.485	.628	E	.99	.319	E	-3.06
26	-.052	-.978	.328	E	.16	.693	E	1.69
27	-.023	-.399	.690	E	.29	.593	E	1.52
28	.061	.854	.393	E	.01	.943	E	-.49
29	.143	1.325	.185	E	.69	.406	E	-1.05
30	.100	1.406	.160	E	.05	.820	E	-.92
31	.253	3.090	.002	E	.74	.391	E	-1.33
32	.119	1.504	.133	E	.27	.605	E	-.98
33	.127	1.246	.213	E	2.83	.093	E	-1.98
34	.102	1.146	.252	E	2.39	.122	E	-2.04
35	-.124	-1.136	.256	E	4.71	.030	E	2.44
36	-.210	-2.109	.035	E	1.91	.166	E	1.61
37	.117	1.211	.226	E	.76	.383	E	-1.11
38	-.042	-1.266	.206	E	.43	.513	E	1.63
39	-.060	-.562	.574	E	.01	.922	E	.13
40	.026	.432	.666	E	.00	.946	E	.51
41	.009	.138	.890	E	.00	.982	E	.61
42	-.148	-1.488	.137	E	1.82	.177	E	1.62

43	-.058	-1.083	.279 E	.04	.836 E	-.19
44	.011	.125	.901 E	.02	.899 E	.13
45	.098	.951	.342 E	.04	.848 E	-.04
46	.005	.072	.942 E	.66	.417 E	-1.43
47	.097	1.165	.244 E	.38	.539 E	-.87
48	.034	.270	.787 E	.02	.889 E	-.07
49	-.015	-.279	.781 E	.95	.329 E	-2.02
50	.014	.483	.629 E	.16	.691 E	-1.65

Reference group = Controlled-oral-language-focused group
Focal group = Communication-focused group

Second run:

The unflagged items from the first run entered the test.

run no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
1	-.048	-.389	.697 E	.00	.978 E	.16
2	-.057	-.690	.490 E	.25	.617 E	.85
3	-.025	-.515	.606 E	.00	.964 E	.83
5	-.097	-1.422	.155 E	.79	.375 E	2.11
7	.098	1.018	.309 E	2.20	.138 E	-1.83
8	-.122	-1.210	.226 E	.44	.509 E	1.11
9	-.054	-.468	.640 E	.28	.598 E	.78
10	-.024	-.555	.579 E	.01	.928 E	.59
11	.025	.225	.822 E	.00	.982 E	-.19
12	-.018	-.334	.739 E	.00	.956 E	-1.42
13	-.002	-.031	.975 E	.06	.806 E	-1.29
14	-.010	-.115	.908 E	.32	.570 E	-.99
15	-.178	-1.772	.076 E	2.80	.094 E	1.94
16	.115	1.384	.167 E	.54	.462 E	-1.30
17	-.180	-1.710	.087 E	.20	.652 E	.72
18	-.020	-.447	.655 E	.02	.898 E	2.27
19	-.075	-1.070	.284 E	.00	.984 E	.59
20	.081	.770	.442 E	.54	.464 E	-.86
22	-.052	-.404	.686 E	.19	.666 E	-.66
23	.019	.555	.579 E	.00	.962 E	-1.15
24	-.059	-.518	.604 E	.61	.436 E	-1.00
25	.067	2.032	.042 E	3.12	.078 E	-4.97
26	-.088	-1.356	.175 E	1.11	.292 E	3.90
27	-.054	-.765	.445 E	1.67	.196 E	3.08
28	.057	.791	.429 E	.02	.896 E	-.30
29	.070	.785	.433 E	.53	.466 E	-1.07
30	.110	2.142	.032 E	.08	.778 E	-1.07
32	.018	.196	.845 E	.35	.553 E	-1.01
33	.161	1.822	.068 E	1.22	.270 E	-1.41
34	.118	1.392	.164 E	1.39	.239 E	-1.53
37	.029	.345	.730 E	.00	.981 E	-.25
38	-.064	-1.108	.268 E	.92	.338 E	2.78
39	-.165	-2.020	.043 E	3.92	.048 E	2.46
40	-.087	-1.140	.254 E	.87	.352 E	1.60
41	-.001	-.017	.986 E	.49	.482 E	2.08
42	-.151	-1.648	.099 E	2.64	.104 E	2.08
43	-.072	-1.029	.303 E	.85	.356 E	2.49
44	-.044	-.455	.649 E	.31	.576 E	.97
45	-.083	-.993	.321 E	.02	.883 E	.10
46	.049	.622	.534 E	.10	.755 E	-.76
47	.009	.119	.905 E	1.42	.234 E	-1.61
48	.039	.418	.676 E	.06	.814 E	-.45
49	.008	.172	.864 E	.60	.439 E	-1.75
50	.003	.068	.946 E	.26	.612 E	1.88

Reference group = Controlled-oral-language-focused group
Focal group = Communication-focused group

Third run:

Valid subtest items:

1	2	3	5	7	8	9	10	11	12
13	14	16	18	19	20	22	23	24	26
27	28	29	32	34	37	38	40	41	43
44	45	46	47	48	49	50			

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
4	-.158	-2.519	.012 E	3.76	.052 E	4.11
6	-.156	-1.906	.057 E	2.22	.136 E	2.68
21	.017	.488	.625 E	.05	.815 E	.49
31	.164	1.774	.076 E	.45	.504 E	-1.03
36	-.225	-2.276	.023 E	4.91	.027 E	2.60
35	-.127	-1.198	.231 E	1.50	.221 E	1.49
15	-.102	-.942	.346 E	2.58	.108 E	1.93
17	-.058	-.556	.578 E	.70	.401 E	1.18
25	.068	1.472	.141 E	1.12	.290 E	-3.05
30	.098	1.593	.111 E	.14	.708 E	-1.09
33	.149	1.419	.156 E	1.47	.225 E	-1.43
39	-.092	-.962	.336 E	.78	.376 E	1.26
42	-.102	-.900	.368 E	.95	.330 E	1.24

Reference group = Controlled-oral-language-focused group
Focal group = Communication-focused group

Fourth run: Tests of group DIF and DTF

Reference group favored items: 31
Focal group favored items: 4, 36, 6

Suspect subtest items:
31

Valid subtest items:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
32	33	34	35	36	37	38	39	40	41
42	43	44	45	46	47	48	49	50	

proportion of Ref. grp. examinees eliminated = .213
proportion of Focal grp. examinees eliminated = .360

		SIB-uni p-value for	Mantel-Haenszel Results		
	SIB-uni z	DIF against either Ref.	Chi	p-value for DIF against either Ref.	Delta
Beta-uni	statistic	or Foc. grp.	sqr.	or Foc. grp.	(D-DIF)
.253	3.090	.002	.74	.391	-1.332

Suspect subtest items:
4 36 6

Valid subtest items:

1	2	3	5	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22
23	24	25	26	27	28	29	30	31	32
33	34	35	37	38	39	40	41	42	43
44	45	46	47	48	49	50			

proportion of Ref. grp. examinees eliminated = .277
proportion of Focal grp. examinees eliminated = .349

		SIB-uni p-value for	Mantel-Haenszel Results		
	SIB-uni z	DTF against either Ref.	Chi	p-value for DIF against either Ref.	Delta
Beta-uni	statistic	or Foc. grp.	sqr.	or Foc. grp.	(D-DIF)
-.515	-3.659	.000			

Suspect subtest items:

4 36 6 31

Valid subtest items:

1	2	3	5	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22
23	24	25	26	27	28	29	30	32	33
34	35	37	38	39	40	41	42	43	44
45	46	47	48	49	50				

proportion of Ref. grp. examinees eliminated = .213

proportion of Focal grp. examinees eliminated = .256

		SIB-uni p-value for DTF against either Ref. or Foc. grp.	Mantel-Haenszel Results		
	SIB-uni z statistic		Chi sqr.	p-value for DIF against either Ref. or Foc. grp.	Delta (D-DIF)
Beta-uni					
-.251	-1.225	.221			

4. Altaic vs. Indo-European

Reference group = Altaic language group
Focal group = Indo-European language group

First run:

All items were tested individually.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
1	-.303	-3.463	.001 E	17.62	.000 E	4.35
2	.086	1.570	.116 E	.93	.334 E	-1.80
3	.063	1.228	.220 E	.46	.496 E	-2.45
4	-.139	-1.979	.048 E	2.48	.115 E	2.20
5	-.004	-.061	.951 E	.02	.898 E	-.67
6	.126	2.011	.044 E	4.25	.039 E	-3.12
7	.108	1.195	.232 E	1.05	.307 E	-1.18
8	.098	1.161	.246 E	.00	.994 E	-.20
9	.027	.278	.781 E	.03	.858 E	-.33
10	-.017	-.670	.503 E	.00	1.000 E	99.00
11	-.314	-3.562	.000 E	14.61	.000 E	4.01
12	.000	.000	1.000 E	.04	.849 E	1.29
13	-.053	-1.187	.235 E	.48	.490 E	2.47
14	.102	1.308	.191 E	2.24	.135 E	-2.13
15	-.330	-4.334	.000 E	21.01	.000 E	5.90
16	-.071	-.864	.388 E	.05	.825 E	-.01
17	-.103	-1.152	.249 E	.11	.740 E	.52
18	-.016	-.626	.532 E	.24	.628 E	-.95
19	.126	2.031	.042 E	6.84	.009 E	-4.69
20	-.126	-1.450	.147 E	1.41	.236 E	1.30
21	.043	.935	.350 E	.61	.435 E	-1.93
22	.183	1.832	.067 E	1.95	.162 E	-1.42
23	-.004	-.092	.927 E	.01	.930 E	-.87
24	.199	2.179	.029 E	4.87	.027 E	-2.41
25	.012	.277	.782 E	.01	.910 E	-.73
26	.015	.366	.714 E	.56	.456 E	-2.12
27	.067	1.302	.193 E	1.32	.251 E	-2.48
28	.155	2.558	.011 E	3.91	.048 E	-3.38
29	.241	2.453	.014 E	4.31	.038 E	-2.28
30	-.002	-.033	.974 E	.26	.609 E	1.37
31	.063	.667	.505 E	.01	.933 E	-.32
32	.252	4.310	.000 E	12.75	.000 E	-5.29
33	-.205	-2.235	.025 E	6.07	.014 E	2.37
34	-.209	-2.252	.024 E	4.32	.038 E	2.31
35	.140	1.460	.144 E	1.96	.162 E	-1.49
36	-.001	-.008	.993 E	.04	.838 E	.40
37	.116	1.197	.231 E	1.19	.275 E	-1.22
38	.075	1.832	.067 E	3.05	.081 E	-4.69
39	-.055	-.689	.491 E	.00	.950 E	.29
40	-.035	-.712	.477 E	.52	.469 E	1.94
41	-.068	-1.692	.091 E	.13	.722 E	.90

42	-.150	-1.578	.115	E	.24	.624 E	.59
43	-.042	-.731	.465	E	.02	.888 E	-.39
44	-.026	-.407	.684	E	.86	.355 E	1.42
45	-.020	-.219	.826	E	.17	.679 E	.71
46	.099	1.271	.204	E	3.88	.049 E	-3.16
47	.127	1.553	.120	E	.86	.355 E	-1.26
48	-.140	-1.587	.112	E	2.17	.141 E	1.56
49	.050	1.061	.288	E	.26	.608 E	-1.49
50	-.028	-.845	.398	E	1.02	.313 E	4.04

Reference group = Altaic language group
Focal group = Indo-European language group

Second run:

The unflagged items from the first run entered the test.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value		Mantel-Haenszel Chi sqr.	p value		Delta (D-DIF)
2	.070	.964	.335	E	.23	.630	E	-1.02
3	.079	1.665	.096	E	.16	.690	E	.58
4	-.123	-1.793	.073	E	5.18	.023	E	3.39
5	.021	.348	.728	E	.00	.983	E	-.39
7	.105	1.283	.200	E	1.59	.207	E	-1.42
8	.064	.841	.400	E	.01	.905	E	-.30
9	-.008	-.090	.928	E	.01	.939	E	.28
10	-.009	-.506	.613	E	.41	.524	E	-.43
12	.008	.191	.848	E	.07	.797	E	1.40
12	.044	.884	.377	E	.19	.661	E	1.30
14	.108	1.662	.097	E	1.83	.176	E	-2.27
16	-.132	-1.729	.084	E	.04	.848	E	.49
17	-.093	-1.155	.248	E	.35	.554	E	.73
18	.002	.099	.921	E	.01	.943	E	-3.37
20	-.113	-1.159	.246	E	1.34	.248	E	1.35
21	.069	1.919	.055	E	1.14	.287	E	-2.15
22	.160	1.922	.055	E	3.32	.069	E	-1.94
23	.032	.788	.430	E	.00	.966	E	-.57
25	.084	2.104	.035	E	.56	.455	E	-3.20
26	.049	1.340	.180	E	.89	.345	E	-3.48
27	.104	2.879	.004	E	.70	.403	E	-1.98
30	-.037	-.803	.422	E	1.69	.193	E	2.93
31	.077	.984	.325	E	.08	.775	E	.55
35	.200	2.121	.034	E	2.23	.135	E	-1.66
36	.020	.308	.758	E	.06	.805	E	.51
37	.107	1.012	.312	E	.14	.706	E	-.52
38	.140	3.465	.001	E	3.13	.077	E	-4.08
39	-.091	-1.142	.253	E	.00	.948	E	-.18
40	-.080	-1.811	.070	E	2.01	.156	E	3.62
41	-.083	-1.371	.170	E	1.38	.241	E	2.71
42	-.093	-1.165	.244	E	.14	.707	E	.50
43	-.015	-.207	.836	E	.42	.518	E	1.61
44	-.001	-.007	.994	E	1.05	.305	E	1.75
45	-.212	-2.712	.007	E	.03	.857	E	.39
46	.176	2.469	.014	E	.45	.501	E	-1.48
47	.045	.560	.575	E	.25	.620	E	-.79
48	-.240	-2.880	.004	E	1.49	.222	E	1.26
49	.110	1.733	.083	E	.29	.589	E	-1.44
50	-.005	-.118	.906	E	2.86	.091	E	5.41

Reference group = Altaic language group
Focal group = Indo-European language group

Third run:

All flagged items were tested against the valid items which were not flagged from the second run.

Valid subtest items:

2	4	5	7	8	9	10	12	13	14
16	18	20	21	22	23	25	26	30	31
35	36	37	38	39	40	41	42	43	44
45	46	47	48	49	50				

run no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
1	-.319	-3.853	.000 E	13.89	.000 E	3.39
3	.044	1.287	.198 E	.14	.709 E	-1.58
6	.145	2.033	.042 E	5.34	.021 E	-3.86
11	-.237	-2.915	.004 E	11.72	.001 E	3.20
15	-.330	-4.640	.000 E	22.38	.000 E	6.88
17	-.044	-.512	.609 E	.45	.504 E	.73
19	.127	2.030	.042 E	5.57	.018 E	-3.86
24	.133	1.743	.081 E	3.52	.061 E	-2.10
27	.053	1.272	.203 E	.93	.335 E	-1.92
28	.140	2.253	.024 E	3.56	.059 E	-3.28
29	.180	2.038	.042 E	4.06	.044 E	-2.00
32	.328	4.763	.000 E	18.27	.000 E	-8.87
33	-.155	-1.939	.052 E	3.41	.065 E	1.90
34	-.157	-1.796	.072 E	1.80	.180 E	1.32

```
Reference group = Altaic language group
Focal group = Indo-European language group

Fourth run: Tests of suspected items from the third run.

Reference group favored item: 24
Focal group favored item: (33, 24)

-------------------------------------------------------------
 Suspect subtest items:
  24

 Valid subtest items:
   1   2   3   4   5   6   7   8   9  10
  11  12  13  14  15  16  17  18  19  20
  21  22  23  25  26  27  28  29  30  31
  32  33  34  35  36  37  38  39  40  41
  42  43  44  45  46  47  48  49  50
 proportion of Ref. grp. examinees eliminated =    .208
 proportion of Focal grp. examinees eliminated =   .247

                          SIB-uni          Mantel-Haenszel
Results
                    p-value for                    p-value for
           SIB-uni  DIF against                    DIF against
              z     either Ref.          Chi       either Ref.
Delta
 Beta-uni statistic  or Foc. grp.         sqr.    or Foc. grp. (D-
DIF)
    .199     2.179        .029            4.87        .027
-2.409

-------------------------------------------------------------
Suspect subtest items:
  33  34

 Valid subtest items:
   1   2   3   4   5   6   7   8   9  10
  11  12  13  14  15  16  17  18  19  20
  21  22  23  24  25  26  27  28  29  30
  31  32  35  36  37  38  39  40  41  42
  43  44  45  46  47  48  49  50
 proportion of Ref. grp. examinees eliminated =    .264
 proportion of Focal grp. examinees eliminated =   .301

                          SIB-uni          Mantel-Haenszel
Results
                    p-value for                    p-value for
           SIB-uni  DTF against                    DIF against
              z     either Ref.          Chi       either Ref.
Delta
 Beta-uni statistic  or Foc. grp.         sqr.    or Foc. grp. (D-
DIF)
   -.346    -2.549        .011
```

Reference group = Altaic language group
Focal group = Indo-European language group

Fifth run: Tests of group DIF and DTF

Reference group favored items: 6, 19, 24, 28, 29, 32
Focal group favored items: 1, 11, 15, 33, 34

Suspect subtest items:
6 19 24 28 29 32

Valid subtest items:
1 2 3 4 5 7 8 9 10 11
12 13 14 15 16 17 18 20 21 22
23 25 26 27 30 31 33 34 35 36
37 38 39 40 41 42 43 44 45 46
47 48 49 50

proportion of Ref. grp. examinees eliminated = .139
proportion of Focal grp. examinees eliminated = .178

		SIB-uni p-value for	Mantel-Haenszel Results p-value for	
	SIB-uni z	DTF against either Ref.	Chi	DIF against either Ref.	Delta
Beta-uni	statistic	or Foc. grp.	sqr.	or Foc. grp.	(D-DIF)
1.256	6.940	.000			

Suspect subtest items:
1 11 15 33 34

Valid subtest items:
2 3 4 5 6 7 8 9 10 12
13 14 16 17 18 19 20 21 22 23
24 25 26 27 28 29 30 31 32 35
36 37 38 39 40 41 42 43 44 45
46 47 48 49 50

proportion of Ref. grp. examinees eliminated = .153
proportion of Focal grp. examinees eliminated = .329

		SIB-uni p-value for	Mantel-Haenszel Results p-value for	
	SIB-uni z	DTF against either Ref.	Chi	DIF against either Ref.	Delta
Beta-uni	statistic	or Foc. grp.	sqr.	or Foc. grp.	(D-DIF)
-1.505	-5.656	.000			

Suspect subtest items:

1 11 15 6 19 28 29 32 24 33
34

Valid subtest items:

2 3 4 5 7 8 9 10 12 13
14 16 17 18 20 21 22 23 25 26
27 30 31 35 36 37 38 39 40 41
42 43 44 45 46 47 48 49 50

proportion of Ref. grp. examinees eliminated = .181
proportion of Focal grp. examinees eliminated = .164

		SIB-uni	Mantel-Haenszel Results		
		p-value for		p-value for	
	SIB-uni	DTF against		DIF against	
	z	either Ref.	Chi	either Ref.	Delta
Beta-uni	statistic	or Foc. grp.	sqr.	or Foc. grp.	(D-DIF)
-.557	-2.262	.024			

5. Indo-European vs. Sino-Tibetan

Reference group = Indo-European language group
Focal group = Sino-Tibetan language group.

First run:

All items were tested individually.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
1	.418	5.056	.000 E	19.85	.000 E	-4.47
2	.357	4.093	.000 E	13.47	.000 E	-4.65
3	-.044	-.901	.368 E	.16	.688 E	1.76
4	-.013	-.163	.871 E	.13	.723 E	-.57
5	.007	.125	.900 E	.02	.878 E	-.65
6	-.067	-.772	.440 E	2.68	.102 E	2.14
7	-.150	-1.489	.137 E	2.99	.084 E	2.04
8	-.174	-2.134	.033 E	5.60	.018 E	3.07
9	-.018	-.167	.867 E	.43	.514 E	-.79
10	.035	.969	.332 E	.01	.914 E	-1.20
11	-.012	-.135	.893 E	.20	.657 E	-.66
12	.022	.503	.615 E	2.41	.121 E	-3.78
13	.032	.484	.629 E	.00	.974 E	-.48
14	.091	1.168	.243 E	1.49	.222 E	-1.45
15	.417	4.990	.000 E	18.34	.000 E	-4.82
16	-.101	-1.211	.226 E	2.89	.089 E	2.36
17	-.046	-.499	.618 E	.00	.978 E	.23
18	.015	.710	.478 E	.16	.687 E	2.42
19	-.157	-2.109	.035 E	1.46	.227 E	2.05
20	.305	3.724	.000 E	8.41	.004 E	-3.02
21	-.033	-.721	.471 E	.02	.880 E	-.48
22	-.354	-3.936	.000 E	12.42	.000 E	4.07
23	-.046	-1.208	.227 E	.21	.647 E	.28
24	-.053	-.562	.574 E	.26	.612 E	.75
25	.081	1.305	.192 E	.05	.818 E	-.76
26	-.018	-.359	.720 E	.03	.852 E	1.08
27	-.123	-2.223	.026 E	2.69	.101 E	3.43
28	-.180	-2.891	.004 E	4.39	.036 E	4.04
29	-.296	-2.734	.006 E	6.03	.014 E	2.55
30	.100	1.470	.141 E	.85	.357 E	-1.76
31	.035	.356	.722 E	.53	.467 E	-.90
32	-.097	-1.094	.274 E	1.48	.223 E	1.55
33	.028	.253	.800 E	.00	.967 E	.14
34	.055	.570	.569 E	.04	.850 E	-.03
35	-.107	-1.072	.284 E	.47	.493 E	.88
36	-.096	-1.056	.291 E	3.04	.081 E	2.17
37	.010	.084	.933 E	.41	.524 E	.80
38	-.101	-2.349	.019 E	1.23	.268 E	3.11
39	-.041	-.417	.676 E	.48	.489 E	.99
40	.198	3.084	.002 E	6.65	.010 E	-4.02
41	.110	1.873	.061 E	.08	.778 E	-.95
42	-.118	-1.345	.179 E	.19	.662 E	.66

43	-.076	-1.194	.232 E	2.19	.139 E	3.04
44	.272	2.962	.003 E	16.98	.000 E	-5.66
45	.044	.504	.614 E	.00	.996 E	-.21
46	-.106	-1.394	.163 E	2.88	.090 E	2.34
47	-.241	-3.309	.001 E	3.00	.083 E	1.96
48	.037	.407	.684 E	.09	.758 E	-.45
49	-.052	-.907	.364 E	.91	.341 E	1.97
50	.044	1.516	.129 E	2.11	.146 E	-5.34

Reference group = Indo-European language group
Focal group = Sino-Tibetan language group.

Second run:

The unflagged items from the first run entered the test.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
3	-.065	-1.537	.124 E	2.03	.155 E	4.36
4	.090	1.048	.295 E	.62	.431 E	-1.11
5	.030	.410	.682 E	.12	.728 E	-1.02
6	-.012	-.163	.870 E	.88	.347 E	1.60
7	-.127	-1.520	.129 E	2.56	.110 E	1.96
9	.017	.167	.867 E	.03	.862 E	-.02
10	.052	1.301	.193 E	.92	.338 E	-4.59
11	-.037	-.444	.657 E	.00	.978 E	-.26
12	.060	1.224	.221 E	3.13	.077 E	-4.65
13	.054	1.040	.298 E	3.01	.083 E	-4.69
14	.166	2.080	.037 E	.79	.374 E	-1.14
16	-.030	-.424	.672 E	1.68	.195 E	1.94
17	-.017	-.160	.873 E	.00	.997 E	.17
18	.000	.005	.996 E	.32	.572 E	.63
19	-.031	-.477	.633 E	.06	.803 E	.80
21	-.017	-.373	.709 E	.11	.742 E	1.49
23	-.004	-.122	.903 E	.12	.732 E	-2.82
24	-.151	-1.473	.141 E	1.85	.173 E	1.46
25	.102	2.421	.015 E	2.70	.100 E	-4.31
26	.008	.171	.864 E	.10	.751 E	.20
27	-.028	-.487	.626 E	1.19	.275 E	2.65
30	.080	1.345	.179 E	1.58	.209 E	-2.21
31	.102	1.203	.229 E	.37	.543 E	-.79
32	-.048	-.559	.576 E	.01	.943 E	.31
33	.035	.365	.715 E	.71	.398 E	-.99
34	.225	2.400	.016 E	.51	.475 E	-.85
35	-.056	-.643	.520 E	.30	.582 E	.77
36	-.116	-1.264	.206 E	2.68	.102 E	1.95
37	-.022	-.197	.844 E	.00	.955 E	-.13
38	-.018	-.417	.676 E	.03	.866 E	1.99
39	.035	.401	.688 E	.00	.990 E	-.29
41	.083	1.474	.141 E	.66	.417 E	-1.55
42	.013	.133	.894 E	.26	.613 E	.72
43	.000	-.008	.994 E	.95	.331 E	3.19
45	.118	1.250	.211 E	.25	.618 E	-.71
46	-.066	-1.015	.310 E	2.56	.110 E	2.28
48	.093	.935	.350 E	.97	.325 E	-1.12
49	-.034	-.629	.529 E	.61	.436 E	2.42
50	.050	1.080	.280 E	1.66	.197 E	*****

Reference group = Indo-European language group
Focal group = Sino-Tibetan language group.

Third run:

All flagged items were tested against the valid items which were not flagged from the second run.

run no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Delta (D-DIF)
1	.372	4.159	.000 E	19.05	.000 E	-4.64
2	.284	3.140	.002 E	15.10	.000 E	-4.50
8	-.119	-1.415	.157 E	1.47	.225 E	1.47
15	.452	6.370	.000 E	27.24	.000 E	-7.25
20	.293	3.232	.001 E	9.82	.002 E	-3.60
22	-.347	-4.073	.000 E	10.99	.001 E	3.23
28	-.156	-2.347	.019 E	4.99	.026 E	4.08
29	-.116	-1.189	.235 E	1.98	.160 E	1.44
40	.226	3.565	.000 E	13.29	.000 E	-6.17
44	.298	3.619	.000 E	12.06	.001 E	-4.05
47	-.125	-1.333	.182 E	2.58	.108 E	2.07
12	.059	.922	.357 E	2.22	.137 E	-3.70
13	.020	.304	.761 E	2.25	.133 E	-3.31
14	.188	2.044	.041 E	1.45	.229 E	-1.52
25	-.013	-.228	.819 E	1.51	.219 E	-3.18
34	.042	.462	.644 E	.95	.331 E	-1.16

Reference group = Indo-European language group
Focal group = Sino-Tibetan language group.

Fourth run: Tests of group DIF and DTF

Reference group favored items: 1, 2, 14, 15, 20, 40, 44
Focal group favored items: 22, 28

Suspect subtest items:
1 2 15 14 20 40 44

Valid subtest items:

3	4	5	6	7	8	9	10	11	12
13	16	17	18	19	21	22	23	24	25
26	27	28	29	30	31	32	33	34	35
36	37	38	39	41	42	43	45	46	47
48	49	50							

proportion of Ref. grp. examinees eliminated = .301
proportion of Focal grp. examinees eliminated = .172

		SIB-uni p-value for DTF against either Ref. or Foc. grp.	Mantel-Haenszel Results		
	SIB-uni z statistic		Chi sqr.	p-value for DIF against either Ref. or Foc. grp.	Delta (D-DIF)
Beta-uni					
2.598	11.097	.000			

Suspect subtest items:
22 28

Valid subtest items:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	23	24	25	26	27	29	30	31	32
33	34	35	36	37	38	39	40	41	42
43	44	45	46	47	48	49	50		

proportion of Ref. grp. examinees eliminated = .288
proportion of Focal grp. examinees eliminated = .172

		SIB-uni p-value for DTF against either Ref. or Foc. grp.	Mantel-Haenszel Results		
	SIB-uni z statistic		Chi sqr.	p-value for DIF against either Ref. or Foc. grp.	Delta (D-DIF)
Beta-uni					
-.550	-5.013	.000			

OUTPUT FOR RUN NUMBER 3 OUTPUT FOR RUN NUMBER 3

Suspect subtest items:
1 2 15 14 20 40 44 22 28

Valid subtest items:
3 4 5 6 7 8 9 10 11 12
13 16 17 18 19 21 23 24 25 26
27 29 30 31 32 33 34 35 36 37
38 39 41 42 43 45 46 47 48 49
50

proportion of Ref. grp. examinees eliminated = .329
proportion of Focal grp. examinees eliminated = .156

		SIB-uni p-value for	Mantel-Haenszel Results		
	SIB-uni z	DTF against either Ref.		p-value for DIF against either Ref.	Delta
Beta-uni	statistic	or Foc. grp.	Chi sqr.	or Foc. grp.	(D-DIF)
1.811	7.190	.000			

6. Sino-Tibetan vs. Altaic

Reference group = Sino-Tibetan language group
Focal group = Altaic language group.

First run:

All items are tested individually.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
1	-.089	-1.054	.292 E	.05	.822 E	.42
2	-.349	-4.296	.000 E	17.07	.000 E	5.31
3	-.007	-.209	.835 E	.01	.927 E	-1.80
4	.100	1.324	.185 E	1.75	.186 E	-2.15
5	-.052	-.972	.331 E	.00	.990 E	.58
6	-.082	-1.419	.156 E	1.12	.290 E	2.51
7	.116	1.109	.267 E	1.13	.288 E	-1.30
8	.147	1.786	.074 E	2.95	.086 E	-2.13
9	.007	.078	.938 E	.00	.947 E	-.13
10	-.049	-1.935	.053 E	.25	.620 E	2.57
11	.333	3.587	.000 E	9.67	.002 E	-3.04
12	-.089	-1.844	.065 E	.77	.381 E	2.60
13	-.070	-1.563	.118 E	.00	.992 E	.63
14	-.215	-2.802	.005 E	5.00	.025 E	3.01
15	-.046	-.482	.630 E	.19	.664 E	.62
16	.062	.758	.448 E	2.59	.107 E	-2.30
17	.135	1.565	.118 E	.44	.506 E	-.85
18	.004	.142	.887 E	.00	1.000 E	-3.78
19	-.035	-.693	.489 E	.36	.549 E	1.65
20	-.122	-1.230	.219 E	1.29	.256 E	1.23
21	.021	.721	.471 E	.32	.570 E	*****
22	.174	2.096	.036 E	6.13	.013 E	-3.58
23	.033	1.206	.228 E	.03	.864 E	-1.99
24	-.107	-1.032	.302 E	.79	.373 E	1.05
25	-.079	-1.673	.094 E	2.42	.120 E	4.17
26	-.047	-1.105	.269 E	.14	.712 E	1.90
27	-.063	-2.129	.033 E	.07	.796 E	2.15
28	.045	1.318	.187 E	.00	.971 E	-.84
29	.055	.515	.607 E	.06	.807 E	-.44
30	-.115	-1.810	.070 E	.00	.969 E	.50
31	-.052	-.618	.537 E	.16	.693 E	.70
32	-.209	-3.107	.002 E	7.30	.007 E	4.26
33	.216	2.246	.025 E	1.74	.187 E	-1.69
34	.095	.986	.324 E	.50	.480 E	-.80
35	-.068	-.761	.447 E	.02	.888 E	-.08
36	.132	1.864	.062 E	3.00	.083 E	-2.04
37	-.111	-1.005	.315 E	.04	.851 E	.36
38	-.005	-.199	.842 E	.03	.870 E	1.83
39	.009	.088	.930 E	.00	.998 E	-.27
40	-.136	-1.931	.053 E	1.45	.228 E	2.12
41	-.040	-.652	.514 E	.00	.989 E	.56

42	.208	2.110	.035	E	1.22	.268 E -1.34
43	.087	1.868	.062	E	2.98	.084 E -4.98
44	-.314	-3.402	.001	E	14.22	.000 E 4.69
45	-.025	-.259	.796	E	.01	.912 E -.11
46	-.065	-1.330	.184	E	.00	.969 E -.88
47	.107	1.402	.161	E	.33	.563 E -1.06
48	.164	1.807	.071	E	.01	.942 E -.26
49	-.017	-.560	.575	E	.04	.841 E -.62
50	-.044	-1.068	.286	E	.00	.964 E .87

Reference group = Sino-Tibetan language group
Focal group = Altaic language group.

Second run:

The unflagged items from the first run entered the test.

Item no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value		Mantel-Haenszel Chi sqr.	p value		Delta (D-DIF)
1	.029	.281	.778	E	.52	.471	E	.87
3	.012	.292	.770	E	.22	.639	E	-.10
4	.124	1.569	.117	E	.21	.646	E	-.85
5	-.010	-.160	.873	E	1.08	.298	E	2.16
6	-.062	-1.210	.226	E	.57	.449	E	1.90
7	.020	.222	.825	E	.14	.708	E	-.56
8	.178	2.340	.019	E	2.68	.102	E	-2.15
9	-.011	-.110	.912	E	.02	.891	E	.05
10	-.059	-1.802	.072	E	.57	.450	E	3.62
12	-.056	-1.034	.301	E	.79	.374	E	2.73
13	-.041	-.763	.446	E	.03	.872	E	.29
15	-.086	-.991	.322	E	.90	.344	E	1.05
16	.096	1.060	.289	E	4.95	.026	E	-2.62
17	.165	1.746	.081	E	1.98	.159	E	-1.47
18	.000	-.003	.997	E	.23	.635	E	.75
19	-.039	-.675	.500	E	.17	.684	E	1.22
20	-.108	-1.137	.256	E	2.77	.096	E	1.85
21	.006	.176	.861	E	.05	.815	E	.36
23	.046	1.294	.196	E	.00	.975	E	-1.44
24	-.103	-1.151	.250	E	.43	.510	E	.92
25	-.077	-1.705	.088	E	.53	.466	E	2.90
26	-.031	-.737	.461	E	.00	.991	E	.91
27	-.057	-1.935	.053	E	.00	.944	E	.83
28	.014	.392	.695	E	.18	.668	E	-.02
29	-.001	-.015	.988	E	.00	.947	E	.23
30	-.034	-.486	.627	E	.19	.659	E	1.11
31	-.049	-.605	.545	E	.98	.321	E	1.38
33	.183	2.139	.032	E	4.88	.027	E	-2.28
34	.068	.755	.450	E	.17	.676	E	-.62
35	-.091	-1.010	.312	E	.00	.965	E	.24
36	.072	.780	.436	E	2.03	.155	E	-1.56
37	-.010	-.104	.917	E	.03	.854	E	-.01
38	-.003	-.090	.929	E	.48	.490	E	-.07
39	-.010	-.115	.908	E	.02	.895	E	-.40
40	-.112	-1.462	.144	E	5.85	.016	E	3.78
41	-.024	-.464	.642	E	.14	.708	E	1.11
42	.095	1.019	.308	E	.83	.361	E	-1.06
43	.067	1.188	.235	E	.79	.373	E	-2.57
45	-.009	-.105	.917	E	.01	.940	E	.27
46	-.007	-.146	.884	E	.25	.617	E	1.49
47	.099	1.368	.171	E	.95	.330	E	-1.30
48	.123	1.219	.223	E	.37	.545	E	-.73
49	-.036	-.836	.403	E	.05	.826	E	1.18
50	-.031	-.756	.449	E	.17	.682	E	1.56

Reference group = Sino-Tibetan language group
Focal group = Altaic language group.

Third run:

All flagged items were tested against the valid items which were not flagged from the second run.

Valid subtest items:

1	3	4	5	6	7	9	12	13	15
16	18	19	20	21	23	24	26	28	29
30	31	34	35	36	37	38	39	41	42
43	45	46	47	48	49	50			

run no.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Mantel-Haenszel Chi sqr.	Mantel-Haenszel p value	Mantel-Haenszel Delta (D-DIF)
2	-.379	-5.040	.000 E	24.66	.000 E	7.76
11	.360	4.062	.000 E	17.05	.000 E	-4.17
14	-.229	-2.945	.003 E	7.38	.007 E	3.41
22	.214	2.763	.006 E	7.28	.007 E	-3.23
32	-.192	-2.935	.003 E	8.21	.004 E	4.67
44	-.276	-3.240	.001 E	10.21	.001 E	3.56
8	.168	2.157	.031 E	1.90	.168 E	-1.81
33	.125	1.340	.180 E	3.94	.047 E	-2.09
40	-.118	-1.516	.129 E	4.29	.038 E	2.52
10	-.058	-1.531	.126 E	.51	.475 E	3.30
17	.140	1.424	.154 E	2.42	.120 E	-1.70
25	-.057	-1.232	.218 E	1.44	.230 E	3.14
27	-.032	-.912	.362 E	.01	.935 E	-1.14

```
Reference group = Sino-Tibetan language group
Focal group = Altaic language group.

Fourth run: Tests of group DIF and DTF

Reference group favored items: 8, 11, 22
Focal group favored items: 2, 14, 32, 44

------------------------------------------------------------

 Suspect subtest items:
   8  11  22

 Valid subtest items:
   1   2   3   4   5   6   7   9  10  12
  13  14  15  16  17  18  19  20  21  23
  24  25  26  27  28  29  30  31  32  33
  34  35  36  37  38  39  40  41  42  43
  44  45  46  47  48  49  50
 proportion of Ref. grp. examinees eliminated =    .250
 proportion of Focal grp. examinees eliminated =   .181

                          SIB-uni                  Mantel-Haenszel Results
                        p-value for                      p-value for
              SIB-uni   DTF against                      DIF against
                 z      either Ref.          Chi        either Ref.    Delta
 Beta-uni statistic  or Foc. grp.           sqr.   or Foc. grp. (D-DIF)
    .914     7.589        .000

------------------------------------------------------------

 Suspect subtest items:
   2  14  32  44

 Valid subtest items:
   1   3   4   5   6   7   8   9  10  11
  12  13  15  16  17  18  19  20  21  22
  23  24  25  26  27  28  29  30  31  33
  34  35  36  37  38  39  40  41  42  43
  45  46  47  48  49  50
 proportion of Ref. grp. examinees eliminated =    .109
 proportion of Focal grp. examinees eliminated =   .208

                          SIB-uni                  Mantel-Haenszel Results
                        p-value for                      p-value for
              SIB-uni   DTF against                      DIF against
                 z      either Ref.          Chi        either Ref.    Delta
 Beta-uni statistic  or Foc. grp.           sqr.   or Foc. grp. (D-DIF)
  -1.193    -8.126        .000
```

Suspect subtest items:

2	11	22	8	14	32	44			

Valid subtest items:

1	3	4	5	6	7	9	10	12	13
15	16	17	18	19	20	21	23	24	25
26	27	28	29	30	31	33	34	35	36
37	38	39	40	41	42	43	45	46	47
48	49	50							

proportion of Ref. grp. examinees eliminated = .078
proportion of Focal grp. examinees eliminated = .125

		SIB-uni p-value for DTF against either Ref.	Mantel-Haenszel Results		
	SIB-uni z			p-value for DIF against either Ref.	
Beta-uni	statistic	or Foc. grp.	Chi sqr.	or Foc. grp.	Delta (D-DIF)
-.275	-1.485	.138			